

---

## Modeling Univariate Distributions

### 5.1 Introduction

As seen in Chapter 4, usually the marginal distributions of financial time series are not well fit by normal distributions. Fortunately, there are a number of suitable alternative models, such as  $t$ -distributions, generalized error distributions, and skewed versions of  $t$ - and generalized error distributions. All of these will be introduced in this chapter. Typically, the parameters in these distributions are estimated by maximum likelihood. Sections 5.9 and 5.14 provide an introduction to the maximum likelihood estimator (MLE), and Section 5.18 provides references for further study on this topic.

Software for maximum likelihood is readily available for standard models, and a reader interested only in data analysis and modeling often need not be greatly concerned with the technical details of maximum likelihood. However, when performing a statistical analysis, it is always worthwhile to understand the underlying theory, at least at a conceptual level, since doing so can prevent misapplications. Moreover, when using a nonstandard model, often there is no software available for automatic computation of the MLE and one needs to understand enough theory to write a program to compute the MLE.

### 5.2 Parametric Models and Parsimony

In a parametric statistical model, the distribution of the data is completely specified except for a finite number of unknown parameters. For example, assume that  $Y_1, \dots, Y_n$  are i.i.d. from a  $t$ -distribution<sup>1</sup> with mean  $\mu$ , variance  $\sigma^2$ , and degrees of freedom  $\nu$ . Then this is a parametric model provided that, as is usually the case, one or more of  $\mu$ ,  $\sigma^2$ , and  $\nu$  are unknown.

---

<sup>1</sup> The reader who is unfamiliar with  $t$ -distributions should look ahead to Section 5.5.2.

A model should have only as many parameters as needed to capture the important features of the data. Each unknown parameter is another quantity to estimate and another source of estimation error. Estimation error, among other things, increases the uncertainty when one forecasts future observations. On the other hand, a statistical model must have enough parameters to adequately describe the behavior of the data. A model with too few parameters can create biases because the model does not fit the data well.

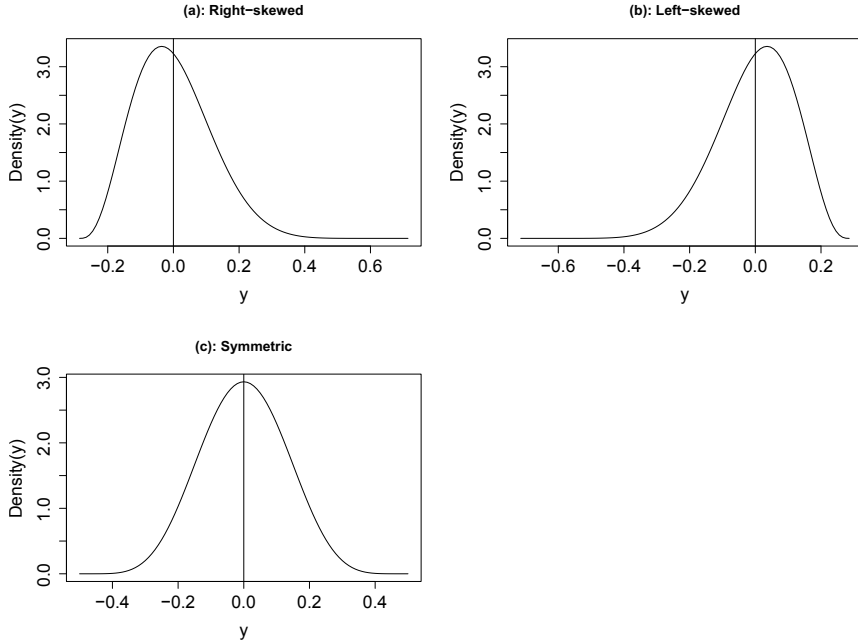
A statistical model with little bias, but without excess parameters, is called *parsimonious* and achieves a good tradeoff between bias and variance. Finding one or a few parsimonious models is an important part of data analysis.

### 5.3 Location, Scale, and Shape Parameters

Parameters are often classified as location, scale, or shape parameters depending upon which properties of a distribution they determine. A *location parameter* is a parameter that shifts a distribution to the right or left without changing the distribution's shape or variability. Scale parameters quantify dispersion. A parameter is a *scale parameter* for a univariate sample if the parameter is increased by the amount  $|a|$  when the data are multiplied by  $a$ . Thus, if  $\sigma(X)$  is a scale parameter for a random variable  $X$ , then  $\sigma(aX) = |a|\sigma(X)$ . A scale parameter is a constant multiple of the standard deviation provided that the latter is finite. Many examples of location and scale parameters can be found in the following sections. If  $\lambda$  is a scale parameter, then  $\lambda^{-1}$  is called an inverse-scale parameter. Since scale parameters quantify dispersion, inverse-scale parameters quantify precision.

If  $f(y)$  is any fixed density, then  $f(y - \mu)$  is a family of distributions with location parameter  $\mu$ ;  $\theta^{-1}f(y/\theta)$ ,  $\theta > 0$ , is a family of distributions with a scale parameter  $\theta$ ; and  $\theta^{-1}f\{\theta^{-1}(y - \mu)\}$  is a family of distributions with location parameter  $\mu$  and scale parameter  $\theta$ . These facts can be derived by noting that if  $Y$  has density  $f(y)$  and  $\theta > 0$ , then, by Result A.6.1,  $Y + \mu$  has density  $f(y - \mu)$ ,  $\theta Y$  has density  $\theta^{-1}f(\theta^{-1}y)$ , and  $\theta Y + \mu$  has density  $\theta^{-1}f\{\theta^{-1}(y - \mu)\}$ .

A *shape* parameter is defined as any parameter that is not changed by location and scale changes. More precisely, for any  $f(y)$ ,  $\mu$ , and  $\theta > 0$ , the value of a shape parameter for the density  $f(y)$  will equal the value of that shape parameter for  $\theta^{-1}f\{\theta^{-1}(y - \mu)\}$ . The degrees-of-freedom parameter for  $t$ -distributions is a shape parameter. Other shape parameters will be encountered later in this chapter. Shape parameters are often used to specify the skewness or tail weight of a distribution.



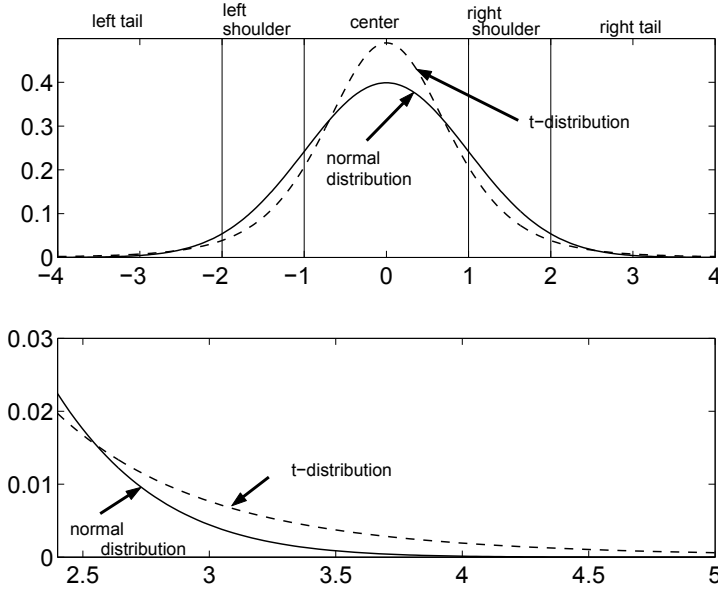
**Fig. 5.1.** *Skewed and symmetric densities. In each case, the mean is zero and is indicated by a vertical line.*

## 5.4 Skewness, Kurtosis, and Moments

Skewness and kurtosis help characterize the shape of a probability distribution. *Skewness* measures the degree of asymmetry, with symmetry implying zero skewness, positive skewness indicating a relatively long right tail compared to the left tail, and negative skewness indicating the opposite. Figure 5.1 shows three densities, all with an expectation equal to 0. The densities are right-skewed, left-skewed, and symmetric about 0, respectively, in panels (a)–(c).

*Kurtosis* indicates the extent to which probability is concentrated in the center and especially the tails of the distribution rather than in the “shoulders,” which are the regions between the center and the tails.

In Section 4.3.2, the left tail was defined as the region from  $-\infty$  to  $\mu - 2\sigma$  and the right tail as the region from  $\mu + 2\sigma$  to  $+\infty$ . Here  $\mu$  and  $\sigma$  could be the mean and standard deviation or the median and MAD. Admittedly, these definitions are somewhat arbitrary. Reasonable definitions of *center* and *shoulder* would be that the center is the region from  $\mu - \sigma$  to  $\mu + \sigma$ , the left shoulder is from  $\mu - 2\sigma$  to  $\mu - \sigma$ , and the right shoulder is from  $\mu + \sigma$  to  $\mu + 2\sigma$ . See the upper plot in Figure 5.2. Because skewness and kurtosis measure shape, they do not depend on the values of location and scale parameters.



**Fig. 5.2.** Comparison of a normal density and a  $t$ -density with 5 degrees of freedom. Both densities have mean 0 and standard deviation 1. The upper plot also shows the center, shoulders, and tail regions.

The skewness of a random variable  $Y$  is

$$\text{Sk} = E \left\{ \frac{Y - E(Y)}{\sigma} \right\}^3 = \frac{E\{Y - E(Y)\}^3}{\sigma^3}.$$

To appreciate the meaning of the skewness, it is helpful to look at an example; the binomial distribution is convenient for that purpose. The skewness of the  $\text{Binomial}(n, p)$  distribution is

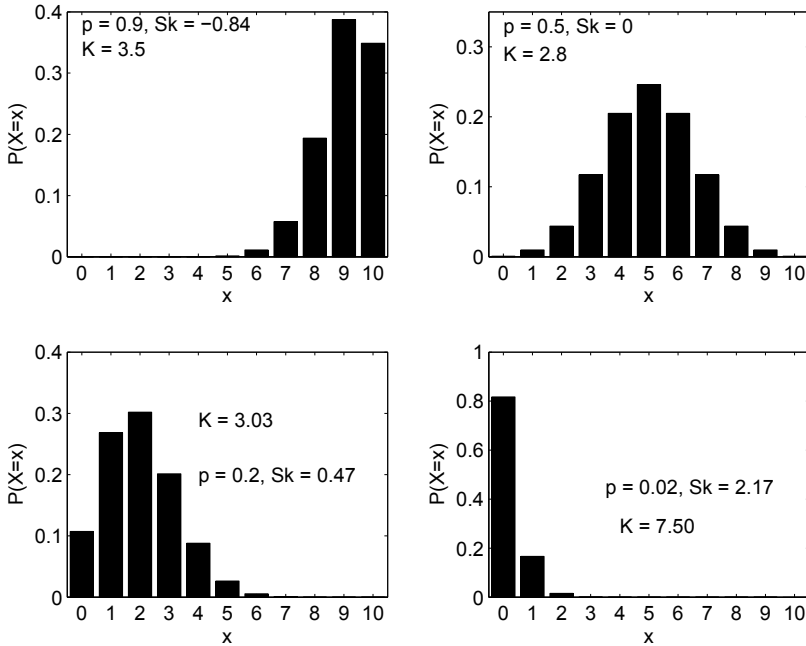
$$\text{Sk}(n, p) = \frac{1 - 2p}{\sqrt{np(1 - p)}}, \quad 0 < p < 1.$$

Figure 5.3 shows the binomial probability distribution and its skewness for  $n = 10$  and four values of  $p$ . Notice that

1. the skewness is positive if  $p < 0.5$ , negative if  $p > 0.5$ , and 0 if  $p = 0.5$ ;
2. the absolute skewness becomes larger as  $p$  moves closer to either 0 or 1 with  $n$  fixed;
3. the absolute skewness decreases to 0 as  $n$  increases to  $\infty$  with  $p$  fixed;

Positive skewness is also called right skewness and negative skewness is called left skewness. A distribution is *symmetric* about a point  $\theta$  if  $P(Y > \theta + y) = P(Y < \theta - y)$  for all  $y > 0$ . In this case,  $\theta$  is a location parameter

and equals  $E(Y)$ , provided that  $E(Y)$  exists. The skewness of any symmetric distribution is 0. Property 3 is not surprising in light of the central limit theorem. We know that the binomial distribution converges to the symmetric normal distribution as  $n \rightarrow \infty$  with  $p$  fixed and not equal to 0 or 1.



**Fig. 5.3.** Several binomial probability distributions with  $n = 10$  and their skewness determined by the shape parameter  $p$ .  $Sk$  = skewness coefficient and  $K$  = kurtosis coefficient. The top left plot has left-skewness ( $Sk = -0.84$ ). The top right plot has no skewness ( $Sk = 0$ ). The bottom left plot has moderate right-skewness ( $Sk = 0.47$ ). The bottom-left plot has strong right skewness ( $Sk = 2.17$ ).

The kurtosis of a random variable  $Y$  is

$$Kur = E \left\{ \frac{Y - E(Y)}{\sigma} \right\}^4 = \frac{E\{Y - E(Y)\}^4}{\sigma^4}.$$

The kurtosis of a normal random variable is 3. The smallest possible value of the kurtosis is 1 and is achieved by any random variable taking exactly two distinct values, each with probability  $1/2$ . The kurtosis of a Binomial( $n, p$ ) distribution is

$$Kur^{\text{Bin}}(n, p) = 3 + \frac{1 - 6p(1 - p)}{np(1 - p)}.$$

Notice that  $\text{Kur}^{\text{Bin}}(n, p) \rightarrow 3$ , the value at the normal distribution, as  $n \rightarrow \infty$  with  $p$  fixed, which is another sign of the central limit theorem at work. Figure 5.3 also gives the kurtosis of the distributions in that figure.  $\text{Kur}^{\text{Bin}}(n, p)$  equals 1, the minimum value of kurtosis, when  $n = 1$  and  $p = 1/2$ .

It is difficult to interpret the kurtosis of an asymmetric distribution because, for such distributions, kurtosis may measure both asymmetry and tail weight, so the binomial is not a particularly good example for understanding kurtosis. For that purpose we will look instead at  $t$ -distributions because they are symmetric. Figure 5.2 compares a normal density with the  $t_5$ -density rescaled to have variance equal to 1. Both have a mean of 0 and a standard deviation of 1. The mean and standard deviation are location and scale parameters, respectively, and do not affect kurtosis. The parameter  $\nu$  of the  $t$ -distribution is a shape parameter. The kurtosis of a  $t_\nu$ -distribution is finite if  $\nu > 4$  and then the kurtosis is

$$\text{Kur}^t(\nu) = 3 + \frac{6}{\nu - 4}. \quad (5.1)$$

For example, the kurtosis is 9 for a  $t_5$ -distribution. Since the densities in Figure 5.2 have the same mean and standard deviation, they also have the same tails, center, and shoulders, at least according to our somewhat arbitrary definitions of these regions, and these regions are indicated on the top plot. The bottom plot zooms in on the right tail. Notice that the  $t_5$ -density has more probability in the tails and center than the  $N(0, 1)$  density. This behavior of  $t_5$  is typical of symmetric distributions with high kurtosis.

Every normal distribution has a skewness coefficient of 0 and a kurtosis of 3. The skewness and kurtosis must be the same for all normal distributions, because the normal distribution has only location and scale parameters, no shape parameters. The kurtosis of 3 agrees with formula (5.1) since a normal distribution is a  $t$ -distribution with  $\nu = \infty$ . The “excess kurtosis” of a distribution is  $(\text{Kur} - 3)$  and measures the deviation of that distribution’s kurtosis from the kurtosis of a normal distribution. From (5.1) we see that the excess kurtosis of a  $t_\nu$ -distribution is  $6/(\nu - 4)$ .

An exponential distribution<sup>2</sup> has a skewness equal to 2 and a kurtosis of 9. A double-exponential distribution has skewness 0 and kurtosis 6. Since the exponential distribution has only a scale parameter and the double-exponential has only a location and a scale parameter, their skewness and kurtosis must be constant.

The Lognormal( $\mu, \sigma^2$ ) distribution, which is discussed in Section A.9.4, has the log-mean  $\mu$  as a scale parameter and the log-standard deviation  $\sigma$  as a shape parameter—even though  $\mu$  and  $\sigma$  are location and scale parameters for the normal distribution itself, they are scale and shape parameters for the lognormal. The effects of  $\sigma$  on lognormal shapes can be seen in Figures 4.11 and A.1. The skewness coefficient of the lognormal( $\mu, \sigma^2$ ) distribution is

<sup>2</sup> The exponential and double-exponential distributions are defined in Section A.9.5.

$$\{\exp(\sigma^2) + 2\}\sqrt{\exp(\sigma^2) - 1}. \quad (5.2)$$

Since  $\mu$  is a scale parameter, it has no effect on the skewness. The skewness increases from 0 to  $\infty$  as  $\sigma$  increases from 0 to  $\infty$ .

Estimation of the skewness and kurtosis of a distribution is relatively straightforward if we have a sample,  $Y_1, \dots, Y_n$ , from that distribution. Let the sample mean and standard deviation be  $\bar{Y}$  and  $s$ . Then the sample skewness, denoted by  $\widehat{\text{Sk}}$ , is

$$\widehat{\text{Sk}} = \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i - \bar{Y}}{s} \right)^3, \quad (5.3)$$

and the sample kurtosis, denoted by  $\widehat{\text{Kur}}$ , is

$$\widehat{\text{Kur}} = \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i - \bar{Y}}{s} \right)^4. \quad (5.4)$$

Often the factor  $1/n$  in (5.3) and (5.4) is replaced by  $1/(n-1)$ . Both the sample skewness and the excess kurtosis should be near 0 if a sample is from a normal distribution. Deviations of the sample skewness and kurtosis from these values are an indication of nonnormality.

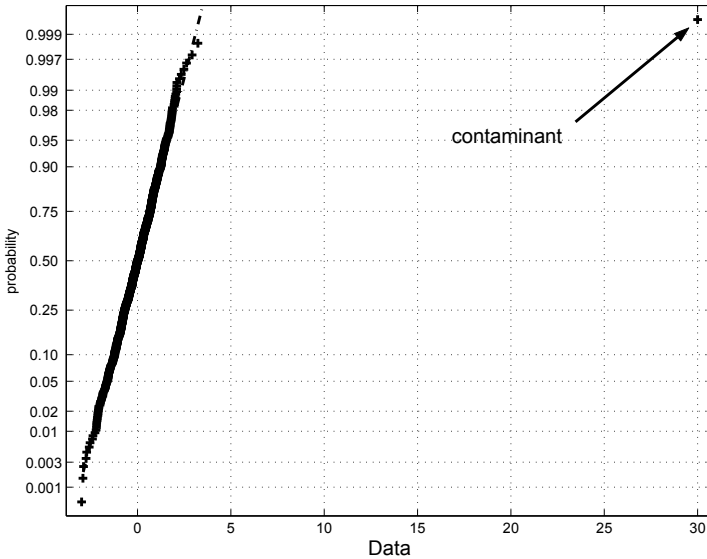


Fig. 5.4. Normal plot of a sample of 999  $N(0, 1)$  data plus a contaminant.

A word of caution is in order. Skewness and kurtosis are highly sensitive to outliers. Sometimes outliers are due to *contaminants*, that is, bad data not from the population being sampled. An example would be a data recording error. A sample from a normal distribution with even a single contaminant that is sufficiently outlying will appear highly nonnormal according to the sample skewness and kurtosis. In such a case, a normal plot *will* look linear, except that the single contaminant will stick out. See Figure 5.4, which is a normal plot of a sample of 999  $N(0, 1)$  data points plus a contaminant equal to 30. This figure shows clearly that the sample is nearly normal but with an outlier. The sample skewness and kurtosis, however, are 10.85 and 243.04, which might give the false impression that the sample is far from normal. Also, even if there were no contaminants, a distribution could be extremely close to a normal distribution and yet have a skewness or excess kurtosis that is very different from 0.

### 5.4.1 The Jarque–Bera test

The Jarque–Bera test of normality compares the sample skewness and kurtosis to 0 and 3, their values under normality. The test statistic is

$$JB = n\{\widehat{Sk}^2/6 + (\widehat{Kur} - 3)^2/24\},$$

which, of course, is 0 when  $\widehat{Sk}$  and  $\widehat{Kur}$ , respectively, have the values 0 and 3, the values expected under normality, and increases as  $\widehat{Sk}$  and  $\widehat{Kur}$  deviate from these values. In R, the test statistic and its  $p$ -value can be computed with the `jarque.bera.test` function.

A large-sample approximation is used to compute a  $p$ -value. Under the null hypothesis, JB converges to the chi-square distribution with 2 degrees of freedom ( $\chi_2^2$ ) as the sample size becomes infinite, so the  $p$ -value is  $1 - F_{\chi_2^2}(JB)$ , where  $F_{\chi_2^2}$  is the CDF of the  $\chi_2^2$ -distribution.

### 5.4.2 Moments

The expectation, variance, skewness coefficient, and kurtosis of a random variable are all special cases of moments, which will be defined in this section.

Let  $X$  be a random variable. The  $k$ th moment of  $X$  is  $E(X^k)$ , so in particular the first moment is the expectation of  $X$ . The  $k$ th absolute moment is  $E|X|^k$ .

The  $k$ th central moment is

$$\mu_k = E[\{X - E(X)\}^k], \quad (5.5)$$

so, for example,  $\mu_2$  is the variance of  $X$ . The skewness coefficient of  $X$  is

$$Sk(X) = \frac{\mu_3}{(\mu_2)^{3/2}}, \quad (5.6)$$



and the kurtosis of  $X$  is

$$\text{Kur}(X) = \frac{\mu_4}{(\mu_2)^2}. \quad (5.7)$$

## 5.5 Heavy-Tailed Distributions

Distributions with higher tail probabilities compared to a normal distribution are called *heavy-tailed*. Because kurtosis is particularly sensitive to tail weight, high kurtosis is nearly synonymous with having a heavy tailed distribution. Heavy-tailed distributions are important models in finance, because equity returns and other changes in market prices usually have heavy tails. In finance applications, one is especially concerned when the return distribution has heavy tails because of the possibility of an extremely large negative return, which could, for example, entirely deplete the capital reserves of a firm. If one sells short,<sup>3</sup> then large positive returns are also worrisome.

### 5.5.1 Exponential and Polynomial Tails

Double-exponential distributions have slightly heavier tails than normal distributions. This fact can be appreciated by comparing their densities. The density of the double-exponential with scale parameter  $\theta$  is proportional to  $\exp(-|y/\theta|)$  and the density of the  $N(0, \sigma^2)$  distribution is proportional to  $\exp\{-0.5(y/\sigma)^2\}$ . The term  $-y^2$  converges to  $-\infty$  much faster than  $-|y|$  as  $|y| \rightarrow \infty$ . Therefore, the normal density converges to 0 much faster than the double-exponential density as  $|y| \rightarrow \infty$ . The generalized error distributions discussed soon in Section 5.6 have densities proportional to

$$\exp(-|y/\theta|^\alpha), \quad (5.8)$$

where  $\alpha > 0$  is a shape parameter and  $\theta$  is a scale parameter. The special cases of  $\alpha = 1$  and 2 are, of course, the double-exponential and normal densities. If  $\alpha < 2$ , then a generalized error distribution will have heavier tails than a normal distribution, with smaller values of  $\alpha$  implying heavier tails. In particular,  $\alpha < 1$  implies a tail heavier than that of a double-exponential distribution.

However, no density of the form (5.8) will have truly heavy tails, and, in particular,  $E(|Y|^k) < \infty$  for all  $k$  so all moments are finite. To achieve a very heavy right tail, the density must be such that

$$f(y) \sim Ay^{-(a+1)} \text{ as } y \rightarrow \infty \quad (5.9)$$

for some  $A > 0$  and  $a > 0$ , which will be called a *right polynomial tail*, rather than like

---

<sup>3</sup> See Section 11.5 for a discussion of short selling.

$$f(y) \sim A \exp(-y/\theta) \text{ as } y \rightarrow \infty \quad (5.10)$$

for some  $A > 0$  and  $\theta > 0$ , which will be called an *exponential right tail*. Polynomial and exponential left tails are defined analogously.

A polynomial tail is also called a *Pareto tail* after the Pareto distribution defined in Section A.9.8. The parameter  $a$  of a polynomial tail is called the *tail index*. The smaller the value of  $a$ , the heavier the tail. The value of  $a$  must be greater than 0, because if  $a \leq 0$ , then the density integrates to  $\infty$ , not 1. An exponential tail as in (5.8) is lighter than any polynomial tail, since

$$\frac{\exp(-|y/\theta|^\alpha)}{|y|^{-(a+1)}} \rightarrow 0 \text{ as } |y| \rightarrow \infty$$

for all  $\theta > 0$ ,  $\alpha > 0$ , and  $a > 0$ .

It is, of course, possible to have left and right tails that behave quite differently from each other. For example, one could be polynomial and the other exponential, or they could both be polynomial but with different indices.

A density with both tails polynomial will have a finite  $k$ th absolute moment only if the smaller of the two tail indices is larger than  $k$ . If both tails are exponential, then all moments are finite.

### 5.5.2 $t$ -Distributions

The  $t$ -distributions have played an extremely important role in classical statistics because of their use in testing and confidence intervals when the data are modeled as having normal distributions. More recently,  $t$ -distributions have gained added importance as models for the distribution of heavy-tailed phenomena such as financial markets data.

We will start with some definitions. If  $Z$  is  $N(0, 1)$ ,  $W$  is chi-squared<sup>4</sup> with  $\nu$  degrees of freedom, and  $Z$  and  $W$  are independent, then the distribution of

$$Z/\sqrt{W/\nu} \quad (5.11)$$

is called the  $t$ -distribution with  $\nu$  *degrees of freedom* and denoted  $t_\nu$ . The  $\alpha$ -upper quantile of the  $t_\nu$ -distribution is denoted by  $t_{\alpha, \nu}$  and is used in tests and confidence intervals about population means, regression coefficients, and parameters in time series models.<sup>5</sup> In testing and interval estimation, the parameter  $\nu$  generally assumes only positive integer values, but when the  $t$ -distribution is used as a model for data,  $\nu$  is restricted only to be positive.

The density of the  $t_\nu$ -distribution is

$$f_{t, \nu}(y) = \left[ \frac{\Gamma\{(\nu+1)/2\}}{(\pi\nu)^{1/2}\Gamma(\nu/2)} \right] \frac{1}{\{1 + (y^2/\nu)\}^{(\nu+1)/2}}. \quad (5.12)$$

Here  $\Gamma$  is the *gamma function* defined by

<sup>4</sup> Chi-squared distributions are discussed in Section A.10.1.

<sup>5</sup> See Section A.17.1 for confidence intervals for the mean.

$$\Gamma(t) = \int_0^\infty x^{t-1} \exp(-x) dx, \quad t > 0. \quad (5.13)$$

The quantity in large square brackets in (5.12) is just a constant, though a somewhat complicated one.

The variance of a  $t_\nu$  is finite and equals  $\nu/(\nu - 2)$  if  $\nu > 2$ . If  $0 < \nu \leq 1$ , then the expected value of the  $t_\nu$ -distribution does not exist and the variance is not defined. If  $1 < \nu \leq 2$ , then the expected value is 0 and the variance is infinite. If  $Y$  has a  $t_\nu$ -distribution, then

$$\mu + \lambda Y$$

is said to have a  $t_\nu(\mu, \lambda^2)$  distribution, and  $\lambda$  will be called *the scale parameter*. With this notation, the  $t_\nu$  and  $t_\nu(0, 1)$  distributions are the same. If  $\nu > 1$ , then the  $t_\nu(\mu, \lambda^2)$  distribution has a mean equal to  $\mu$ , and if  $\nu > 2$ , then it has a variance equal to  $\lambda^2 \nu/(\nu - 2)$ .

The  $t$ -distribution will also be called the *classical  $t$ -distribution* to distinguish it from the standardized  $t$ -distribution defined in the next section.

### Standardized $t$ -Distributions

Instead of the classical  $t$ -distribution just discussed, some software uses a “standardized” version of the  $t$ -distribution. The difference between the two versions is merely notational, but it is important to be aware of this difference.

The  $t_\nu\{0, (\nu - 2)/\nu\}$  distribution with  $\nu > 2$  has a mean equal to 0 and variance equal to 1 and is called a *standardized  $t$ -distribution*, and will be denoted by  $t_\nu^{\text{std}}(0, 1)$ . More generally, for  $\nu > 2$ , define the  $t_\nu^{\text{std}}(\mu, \sigma^2)$  distribution to be equal to the  $t_\nu[\mu, \{(\nu - 2)/\nu\}\sigma^2]$  distribution, so that  $\mu$  and  $\sigma^2$  are the mean and variance of the  $t_\nu^{\text{std}}(\mu, \sigma^2)$  distribution. For  $\nu \leq 2$ ,  $t_\nu^{\text{std}}(\mu, \sigma^2)$  cannot be defined since the  $t$ -distribution does not have a finite variance in this case. The advantage in using the  $t_\nu^{\text{std}}(\mu, \sigma^2)$  distribution is that  $\sigma^2$  is the variance, whereas for the  $t_\nu(\mu, \lambda^2)$  distribution,  $\lambda^2$  is not the variance but instead  $\lambda^2$  is the variance times  $(\nu - 2)/\nu$ .

Some software uses the standardized  $t$ -distribution while other software uses the classical  $t$ -distribution. It is, of course, important to understand which  $t$ -distribution is being used in any specific application. However, estimates from one model can be translated easily into the estimates one would obtain from the other model; see Section 5.14 for an example.

### $t$ -Distributions Have Polynomial Tails

The  $t$ -distributions are a class of heavy-tailed distributions and can be used to model heavy-tail returns data. For  $t$ -distributions, both the kurtosis and the weight of the tails increase as  $\nu$  gets smaller. When  $\nu \leq 4$ , the tail weight is so high that the kurtosis is infinite. For  $\nu > 4$ , the kurtosis is given by (5.1).

By (5.12), the  $t$ -distribution's density is proportional to

$$\frac{1}{\{1 + (y^2/\nu)\}^{(\nu+1)/2}}$$

which for large values of  $|y|$  is approximately

$$\frac{1}{(y^2/\nu)^{(\nu+1)/2}} \propto |y|^{-(\nu+1)}.$$

Therefore, the  $t$ -distribution has polynomial tails with tail index  $a = \nu$ . The smaller the value of  $\nu$ , the heavier the tails.

### 5.5.3 Mixture Models

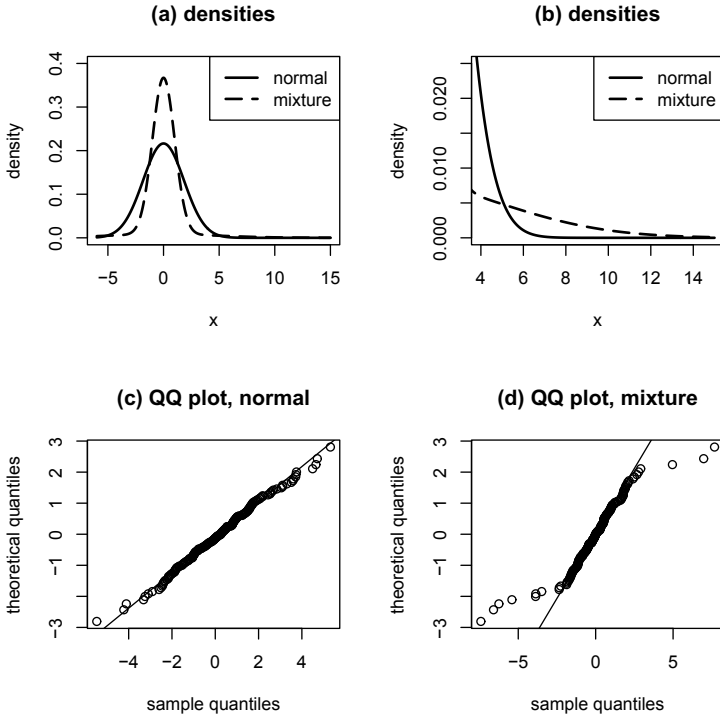
#### Discrete Mixtures

Another class of models containing heavy-tailed distributions is the set of *mixture models*. Consider a distribution that is 90%  $N(0, 1)$  and 10%  $N(0, 25)$ . A random variable  $Y$  with this distribution can be obtained by generating a normal random variable  $X$  with mean 0 and variance 1 and a uniform(0,1) random variable  $U$  that is independent of  $X$ . If  $U < 0.9$ , then  $Y = X$ . If  $U \geq 0.9$ , then  $Y = 5X$ . If an independent sample from this distribution is generated, then the expected percentage of observations from the  $N(0, 1)$  component is 90%. The actual percentage is random; in fact, it has a Binomial( $n, 0.9$ ) distribution, where  $n$  is a sample size. By the law of large numbers, the actual percentage converges to 90% as  $n \rightarrow \infty$ . This distribution could be used to model a market that has two *regimes*, the first being “normal volatility” and second “high volatility,” with the first regime occurring 90% of the time.

This is an example of a *finite* or *discrete normal mixture distribution*, since it is a mixture of a finite number, here two, different normal distributions called the *components*. A random variable with this distribution has a variance equal to 1 with 90% probability and equal to 25 with 10% probability. Therefore, the variance of this distribution is  $(0.9)(1) + (0.1)(25) = 3.4$ , so its standard deviation is  $\sqrt{3.4} = 1.84$ . This distribution is much different than an  $N(0, 3.4)$  distribution, even though the two distributions have the same mean and variance. To appreciate this, look at [Figure 5.5](#).

You can see in [Figure 5.5\(a\)](#) that the two densities look quite different. The normal density looks much more dispersed than the normal mixture, but they actually have the same variances. What is happening? Look at the detail of the right tails in panel (b). The normal mixture density is much higher than the normal density when  $x$  is greater than 6. This is the “outlier” region (along with  $x < -6$ ).<sup>6</sup> The normal mixture has far more outliers than

<sup>6</sup> There is nothing special about “6” to define the boundary of the outlier range, but a specific number was needed to make numerical comparisons. Clearly,  $|x| > 7$  or  $|x| > 8$ , say, would have been just as appropriate as outlier ranges.



**Fig. 5.5.** Comparison of  $N(0, 3.4)$  distribution and heavy-tailed normal mixture distributions. Both distributions have the same mean and variance. The normal mixture distribution is 90%  $N(0, 1)$  and 10%  $N(0, 25)$ . In (c) and (d) the sample size is 200.

the normal distribution and the outliers come from the 10% of the population with a variance of 25. Remember that  $\pm 6$  is only  $6/5$  standard deviations from the mean, using the standard deviation 5 of the component from which they come. Thus, these observations are not outlying relative to their component's standard deviation of 5, only relative to the population standard deviation of  $\sqrt{3.4} = 1.84$  since  $6/1.84 = 3.25$  and three or more standard deviations from the mean is generally considered rather outlying.

Outliers have a powerful effect on the variance and this small fraction of outliers inflates the variance from 1.0 (the variance of 90% of the population) to 3.4.

Let's see how much more probability the normal mixture distribution has in the outlier range  $|x| > 6$  compared to the normal distribution. For an  $N(0, \sigma^2)$  random variable  $Y$ ,

$$P\{|Y| > y\} = 2\{1 - \Phi(y/\sigma)\}.$$

Therefore, for the normal distribution with variance 3.4,

$$P\{|Y| > 6\} = 2\{1 - \Phi(6/\sqrt{3.4})\} = 0.0011.$$

For the normal mixture population that has variance 1 with probability 0.9 and variance 25 with probability 0.1, we have that

$$\begin{aligned} P\{|Y| > 6\} &= 2\left[0.9\{1 - \Phi(6)\} + 0.1\{1 - \Phi(6/5)\}\right] \\ &= 2\{(0.9)(0) + (0.1)(0.115)\} = 0.023. \end{aligned}$$

Since  $0.023/0.0011 \approx 21$ , the normal mixture distribution is 21 times more likely to be in this outlier range than the  $N(0, 3.4)$  population, even though both have a variance of 3.4. In summary, the normal mixture is much more prone to outliers than a normal distribution with the same mean and standard deviation. So, we should be much more concerned about very large negative returns if the return distribution is more like the normal mixture distribution than like a normal distribution. Large positive returns are also likely under a normal mixture distribution and would be of concern when an asset was sold short.

It is not difficult to compute the kurtosis of this normal mixture. Because a normal distribution has kurtosis equal to 3, if  $Z$  is  $N(\mu, \sigma^2)$ , then  $E(Z - \mu)^4 = 3\sigma^4$ . Therefore, if  $Y$  has this normal mixture distribution, then

$$E(Y^4) = 3\{0.9 + (0.1)25^2\} = 190.2$$

and the kurtosis of  $X$  is  $190.2/3.4^2 = 16.45$ .

Normal probability plots of samples of size 200 from the normal and normal mixture distributions are shown in panels (c) and (d) of Figure 5.5. Notice how the outliers in the normal mixture sample give the probability plot a convex-concave pattern typical of heavy-tailed data. The deviation of the plot of the normal sample from linearity is small and is due entirely to randomness.

In this example, the conditional variance of any observations is 1 with probability 0.9 and 25 with probability 0.1. Because there are only two components, the conditional variance is discrete, in fact, with only two possible values, and the example was easy to analyze. This example is a normal *scale mixture* because only the scale parameter  $\sigma$  varies between components. It is also a *discrete mixture* because there are only a finite number of components.

## Continuous Mixtures

The marginal distributions of the GARCH processes studied in Chapter 18 are also normal scale mixtures, but with infinitely many components and a continuous distribution of the conditional variance. Although GARCH processes are more complex than the simple mixture model in this section, the same theme applies—a nonconstant conditional variance of a mixture distribution induces heavy-tailed marginal distributions even though the conditional distributions are normal distributions and have relatively light tails.

The general definition of a normal scale mixture is that it is the distribution of the random variable

$$\mu + \sqrt{U}Z \quad (5.14)$$

where  $\mu$  is a constant equal to the mean,  $Z$  is  $N(0, 1)$ ,  $U$  is a positive random variable giving the variance of each component, and  $Z$  and  $U$  are independent. If  $U$  can assume only a finite number of values, then (5.14) is a *discrete* (or finite) scale mixture distribution. If  $U$  is continuously distributed, then we have a *continuous scale mixture distribution*. The distribution of  $U$  is called the *mixing distribution*. By (5.11), a  $t_\nu$ -distribution is a continuous normal scale mixture with  $\mu = 0$  and  $U = \nu/W$ , where  $\nu$  and  $W$  are as defined above equation (5.11).

Despite the apparent heavy tails of a *finite* normal mixture, the tails are exponential, not polynomial. A continuous normal mixture can have a polynomial tail if the mixture distribution's tail is heavy enough, e.g., as in  $t$ -distributions.

## 5.6 Generalized Error Distributions

Generalized error distributions mentioned briefly in Section 5.5.1 have exponential tails. This section provides more detailed information about them. The standardized generalized error distribution, or GED, with shape parameter  $\nu$  has density

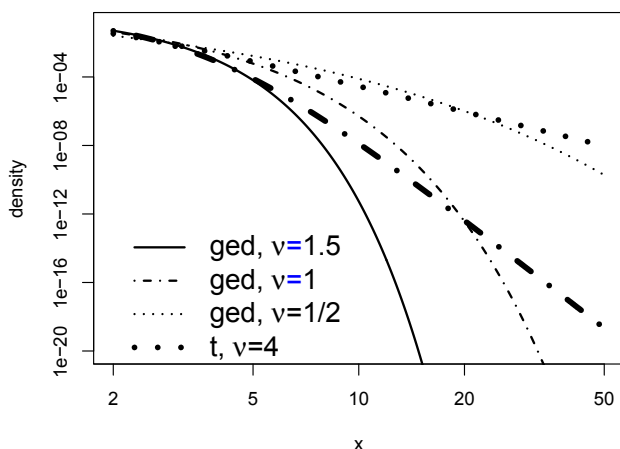
$$f_{\text{ged}}^{\text{std}}(y|\nu) = \kappa(\nu) \exp \left\{ -\frac{1}{2} \left| \frac{y}{\lambda_\nu} \right|^\nu \right\}, \quad -\infty < y < \infty,$$

where  $\kappa(\nu)$  and  $\lambda_\nu$  are constants given by

$$\lambda_\nu = \left\{ \frac{2^{-2/\nu} \Gamma(\nu^{-1})}{\Gamma(3/\nu)} \right\}^{1/2} \quad \text{and} \quad \kappa(\nu) = \frac{\nu}{\lambda_\nu 2^{1+1/\nu} \Gamma(\nu^{-1})}$$

and were chosen so that the function integrates to 1, as it must to be a density, and the variance is 1. The latter property is not necessary but is often convenient.

The shape parameter  $\nu > 0$  determines the tail weight, with smaller values of  $\nu$  giving greater tail weight. When  $\nu = 2$ , a GED is a normal distribution, and when  $\nu = 1$ , it is a double-exponential distribution. The generalized error distributions can give tail weights intermediate between the normal and double-exponential distributions by having  $1 < \nu < 2$ . They can also give tail weights more extreme than the double-exponential distribution by having  $\nu < 1$ .



**Fig. 5.6.** A comparison of the tails of several generalized error (thin curves) and  $t$ -distributions (thick curves).

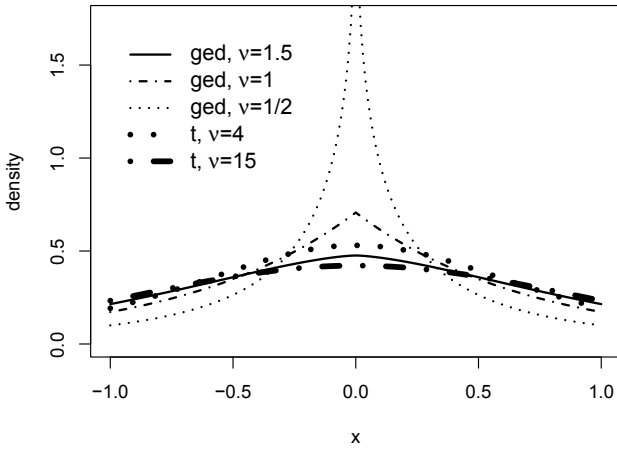
Figure 5.6 shows the right tails of several  $t$ - and generalized error densities with mean 0 and variance 1.<sup>7</sup> Since they are standardized, the argument  $y$  is the number of standard deviations from the median of 0. Because  $t$ -distributions have polynomial tails, any  $t$ -distribution is heavier-tailed than any generalized error distribution. However, this is only an asymptotic result as  $y \rightarrow \infty$ . In the more practical range of  $y$ , tail weight depends as much on the tail weight parameter as it does on the choice between a  $t$ -distribution or a generalized error distribution.

The  $t$ -distributions and generalized error densities also differ in their shapes at the median. This can be seen in Figure 5.7, where the generalized error densities have sharp peaks at the median with the sharpness increasing as  $\nu$  decreases. In comparison, a  $t$ -density is smooth and rounded near the median, even with  $\nu$  small. If a sample is better fit by a  $t$ -distribution than by a generalized error distribution, this may be due more to the sharp central peaks of generalized error densities than to differences between the tails of the two types of distributions.

The  $f_{\text{ged}}^{\text{std}}(y|\nu)$  density is symmetric about 0, which is its mean, median, and mode, and has a variance equal to 1. However, it can be shifted and rescaled to create a location-scale family. The GED distribution with mean  $\mu$ , variance  $\sigma^2$ , and shape parameter  $\nu$  has density

<sup>7</sup> This plot and Figure 5.7 used the R functions `dged` and `dstd` in the `fGarch` package.





**Fig. 5.7.** A comparison of the centers of several generalized error (thin) and  $t$ -densities (thick) with mean 0 and variance 1.

$$f_{\text{ged}}^{\text{std}}(y|\mu, \sigma^2, \nu) := f_{\text{ged}}^{\text{std}}\{(y - \mu)/\sigma|\nu\}/\sigma.$$

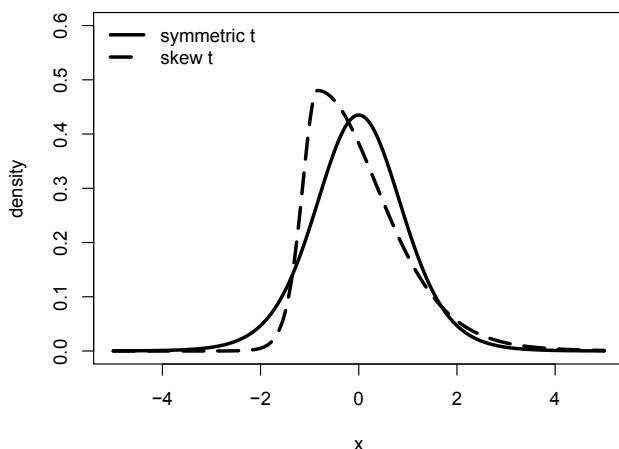
## 5.7 Creating Skewed from Symmetric Distributions

Returns and other financial markets data typically have no natural lower or upper bounds, so one would like to use models with support equal to  $(-\infty, \infty)$ . This is fine if the data are symmetric since then one can use, for example, normal,  $t$ , or generalized error distributions as models. What if the data are skewed? Unfortunately, many of the well-known skewed distributions, such as, gamma and log-normal distributions, have support  $[0, \infty)$  and so are not suitable for many types of financial markets data. This section describes a remedy to this problem.

Fernandez and Steel (1998) have devised a clever way for inducing skewness in symmetric distributions such as normal and  $t$ -distributions. The **fGarch** package in R implements their idea. Let  $\xi$  be a positive constant and  $f$  a density that is symmetric about 0. Define

$$f^*(y|\xi) = \begin{cases} f(y\xi) & \text{if } y < 0, \\ f(y/\xi) & \text{if } y \geq 0. \end{cases} \quad (5.15)$$

Since  $f^*(y|\xi)$  integrates to  $(\xi + \xi^{-1})/2$ ,  $f^*(y|\xi)$  is divided by this constant to create a probability density. After this normalization, the density is given a



**Fig. 5.8.** Symmetric (solid) and skewed (dashed)  $t$ -densities, both with mean 0, standard deviation 1, and  $\nu = 10$ .  $\xi = 2$  in the skewed density. Notice that the mode of the skewed density lies to the left of its mean, a typical behavior of right-skewed densities.

location shift and scale change to induce a mean equal to 0 and variance of 1. The final result is denoted by  $f(y|\xi)$ .

If  $\xi > 1$ , then the right half of  $f(y|\xi)$  is elongated relative to the left hand, which induces right skewness. Similarly,  $\xi < 1$  induces left skewness. Figure 5.8 shows standardized symmetric and skewed  $t$ -distributions<sup>8</sup> with  $\nu = 10$  in both cases and  $\xi = 2$  for the skewed distribution.

If  $f$  is a  $t$ -distribution, then  $f(y|\xi)$  is called a skewed  $t$ -distribution. Skewed  $t$ -distributions include symmetric  $t$ -distributions as special cases where  $\xi = 1$ . In the same way, skewed generalized error distributions are created when  $f$  is a generalized error distribution. The skewed distributions just described will be called Fernandez–Steel or F-N skewed distributions.

Fernandez and Steel’s technique is not the only method for creating skewed versions of the normal and  $t$ -distributions. Azzalini and Capitanio (2003) have created somewhat different skewed normal and  $t$ -distributions.<sup>9</sup> These distributions have a shape parameter  $\alpha$  that determines the skewness; the dis-

<sup>8</sup> R’s `dstd` (for symmetric  $t$ ) and `dsstd` (for skewed  $t$ ) functions in the `fGarch` package were used for to create this plot.

<sup>9</sup> Programs for fitting these distributions, computing their densities, quantile, and distribution functions, and generating random samples are available in R’s `sn` package.

tributed is left-skewed, symmetric, or right-skewed according to whether  $\alpha$  is negative, zero, or positive.

An example is given in Section 5.14 and multivariate versions are discussed in Section 7.9. We will refer to these as Azzalini–Capitanio or A-C skewed distributions.

## 5.8 Quantile-Based Location, Scale, and Shape Parameters

As has been seen, the mean, standard deviation, skewness coefficient, and kurtosis are moments-based location, scale, and shape parameters. Although they are widely used, they have the drawbacks that they are sensitive to outliers and may be undefined or infinite for distributions with heavy tails. An alternative is to use parameters based on quantiles.

Any quantile  $F^{-1}(p)$ ,  $0 < p < 1$ , is a location parameter. A positive weighted average of quantiles, that is,  $\sum_{\ell=1}^L w_{\ell} F^{-1}(p_{\ell})$ , where  $w_{\ell} > 0$  for all  $\ell$  and  $\sum_{\ell=1}^L w_{\ell} = 1$ , is also a location parameter. A simple example is  $\{F^{-1}(1-p) + F^{-1}(p)\}/2$  where  $0 < p < 1/2$ , which equals the mean and median if  $F$  is symmetric.

A scale parameter can be obtained from the difference between two quantiles:

$$s(p_1, p_2) = \frac{F^{-1}(p_2) - F^{-1}(p_1)}{a}$$

where  $0 < p_1 < p_2 < 1$  and  $a$  is a positive constant. An obvious choice is  $p_1 < 1/2$  and  $p_2 = 1 - p_1$ . If  $a = \Phi^{-1}(p_2) - \Phi^{-1}(p_1)$ , then  $s(p_1, p_2)$  is equal to the standard deviation when  $F$  is a normal distribution. If  $a = 1$ , then  $s(1/4, 3/4)$  is called the *interquartile range* or IQR.

A quantile-based shape parameter that quantifies skewness is a ratio with the numerator the difference between two scale parameters and the denominator a scale parameter:

$$\frac{s(1/2, p_2) - s(1/2, p_1)}{s(p_3, p_4)}. \quad (5.16)$$

where  $p_1 < 1/2$ ,  $p_2 > 1/2$ , and  $0 < p_3 < p_4 < 1$ . For example, one could use  $p_2 = 1 - p_1$ ,  $p_4 = p_2$ , and  $p_3 = p_1$ .

A quantile-based shape parameter that quantifies tail weight is the ratio of two scale parameters:

$$\frac{s(p_1, 1 - p_1)}{s(p_2, 1 - p_2)}, \quad (5.17)$$

where  $0 < p_1 < p_2 < 1/2$ . For example, one might have  $p_1 = 0.01$  or  $0.05$  and  $p_2 = 0.25$ .

## 5.9 Maximum Likelihood Estimation

Maximum likelihood is the most important and widespread method of estimation. Many well-known estimators such as the sample mean, and the least-squares estimator in regression are maximum likelihood estimators if the data have a normal distribution. Maximum likelihood estimation generally provides more efficient (less variable) estimators than other techniques of estimation. As an example, for a  $t$ -distribution, the maximum likelihood estimator of the mean is more efficient than the sample mean.

Let  $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$  be a vector of data and let  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$  be a vector of parameters. Let  $f(\mathbf{Y}|\boldsymbol{\theta})$  be the density of  $\mathbf{Y}$ , which depends on the parameters.

The function  $L(\boldsymbol{\theta}) = f(\mathbf{Y}|\boldsymbol{\theta})$  viewed as a function of  $\boldsymbol{\theta}$  with  $\mathbf{Y}$  fixed at the observed data is called the *likelihood function*. It tells us the likelihood of the sample that was actually observed. The *maximum likelihood estimator* (MLE) is the value of  $\boldsymbol{\theta}$  that maximizes the likelihood function. In other words, the MLE is the value of  $\boldsymbol{\theta}$  at which the likelihood of the observed data is largest. We denote the MLE by  $\hat{\boldsymbol{\theta}}_{\text{ML}}$ . Often it is mathematically easier to maximize  $\log\{L(\boldsymbol{\theta})\}$ . If the data are independent, then the likelihood is the product of the marginal densities and products are cumbersome to differentiate. Also, in numerical computations, using the log-likelihood reduces the possibility of underflow or overflow. Taking the logarithm converts the product into an easily differentiated sum. Since the log function is increasing, maximizing  $\log\{L(\boldsymbol{\theta})\}$  is equivalent to maximizing  $L(\boldsymbol{\theta})$ .

In examples found in introductory statistics textbooks, it is possible to find an explicit formula for the MLE. With more complex models such as the ones we will mostly be using, there is no explicit formula for the MLE. Instead, one must write a program that computes  $\log\{L(\boldsymbol{\theta})\}$  for any  $\boldsymbol{\theta}$  and then use optimization software to maximize this function numerically; see Example 5.8. However, for many important models, such as, the examples in the Section 5.14 and the ARIMA and GARCH time series models discussed in Chapter 9, R and other software packages contain functions to find the MLE for these models.

## 5.10 Fisher Information and the Central Limit Theorem for the MLE

Standard errors are essential for gauging the accuracy of estimators. We have formulas for the standard errors of simple estimators such as  $\bar{Y}$ , but what about standard errors for other estimators? Fortunately, there is a simple method for calculating the standard error of a maximum likelihood estimator. We assume for now that  $\boldsymbol{\theta}$  is one-dimensional. The *Fisher information* is defined to be minus the expected second derivative of the log-likelihood, so if  $\mathcal{I}(\boldsymbol{\theta})$  denotes the Fisher information, then

$$\mathcal{I}(\theta) = -E \left[ \frac{d^2}{d\theta^2} \log\{L(\theta)\} \right]. \quad (5.18)$$

The standard error of  $\hat{\theta}$  is simply the inverse square root of the Fisher information, with the unknown  $\theta$  replaced by  $\hat{\theta}$ :

$$s_{\hat{\theta}} = \frac{1}{\sqrt{\mathcal{I}(\hat{\theta})}}. \quad (5.19)$$

*Example 5.1. Fisher information for a normal model mean*

Suppose that  $Y_1, \dots, Y_n$  are i.i.d.  $N(\mu, \sigma^2)$  with  $\sigma^2$  known. The log-likelihood for the unknown parameter  $\mu$  is

$$\log\{L(\mu)\} = -\frac{n}{2}\{\log(\sigma^2) + \log(2\pi)\} - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu)^2.$$

Therefore,

$$\frac{d}{d\mu} \log\{L(\mu)\} = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \mu),$$

and

$$\frac{d^2}{d\mu^2} \log\{L(\mu)\} = -\frac{\sum_{i=1}^n 1}{\sigma^2} = -\frac{n}{\sigma^2}.$$

It follows that  $\mathcal{I}(\hat{\mu}) = n/\sigma^2$  and  $s_{\hat{\mu}} = \sigma/\sqrt{n}$ . Since the MLE is  $\hat{\mu} = \bar{Y}$ , this result is the familiar fact that when  $\sigma$  is known, then  $s_{\bar{Y}} = \sigma/\sqrt{n}$  and when  $\sigma$  is unknown, then  $s_{\bar{Y}} = s/\sqrt{n}$ . □

The theory justifying using these standard errors is the central limit theorem for the maximum likelihood estimator. This theorem can be stated in a mathematically precise manner that is difficult to understand without training in advanced probability theory. The following less precise statement is more easily understood:

**Theorem 5.2.** *Under suitable assumptions, for large enough sample sizes, the maximum likelihood estimator is approximately normally distributed with mean equal to the true parameter and with variance equal to the inverse of the Fisher information.*

The central limit theorem for the maximum likelihood estimator justifies the following large-sample confidence interval for the MLE of  $\theta$ :

$$\hat{\theta} \pm s_{\hat{\theta}} z_{\alpha/2}, \quad (5.20)$$

where  $z_{\alpha/2}$  is the  $\alpha/2$ -upper quantile of the normal distribution and  $s_{\hat{\theta}}$  is defined in (5.19).

The observed Fisher information is

$$\mathcal{I}^{\text{obs}}(\theta) = -\frac{d^2}{d\theta^2} \log\{L(\theta)\}. \quad (5.21)$$

which differs from (5.18) in that there is no expectation taken. In many examples, (5.21) is a sum of many independent terms and, by the law of large numbers, will be close to (5.18). The expectation in (5.18) may be difficult to compute and using (5.21) instead is a convenient alternative.

The standard error of  $\hat{\theta}$  based on observed Fisher information is

$$s_{\hat{\theta}}^{\text{obs}} = \frac{1}{\sqrt{\mathcal{I}^{\text{obs}}(\hat{\theta})}}. \quad (5.22)$$

Often  $s_{\hat{\theta}}^{\text{obs}}$  is used in place of  $s_{\hat{\theta}}$  in the confidence interval (5.20). There is theory suggesting that using the observed Fisher information will result in a more accurate confidence interval, that is, an interval with the true coverage probability closer to the nominal value of  $1 - \alpha$ , so observed Fisher information can be justified by more than mere convenience; see Section 5.18.

So far, it has been assumed that  $\theta$  is one-dimensional. In the multivariate case, the second derivative in (5.18) is replaced by the Hessian matrix of second derivatives, and the result is called the *Fisher information matrix*. Analogously, the observed Fisher information matrix is the multivariate analog of (5.21). Fisher information matrices are discussed in more detail in Section 7.10.

## Bias and Standard Deviation of the MLE

In many examples, the MLE has a small bias that decreases to 0 at rate  $n^{-1}$  as the sample size  $n$  increases to  $\infty$ . More precisely,

$$\text{BIAS}(\hat{\theta}_{\text{ML}}) = E(\hat{\theta}_{\text{ML}}) - \theta \sim \frac{A}{n}, \text{ as } n \rightarrow \infty, \quad (5.23)$$

for some constant  $A$ . The bias of the MLE of a normal variance is an example and  $A = -\sigma^2$  in this case.

Although this bias can be corrected in some special problems, such as, estimation of a normal variance, usually the bias is ignored. There are two good reasons for this. First, the log-likelihood usually is the sum of  $n$  terms and so grows at rate  $n$ . The same is true of the Fisher information. Therefore, the variance of the MLE decreases at rate  $n^{-1}$ , that is,

$$\text{Var}(\hat{\theta}_{\text{ML}}) \sim \frac{B}{n}, \text{ as } n \rightarrow \infty, \quad (5.24)$$

for some  $B > 0$ . Variability should be measured by the standard deviation, not the variance, and by (5.24),

$$\text{SD}(\hat{\theta}_{\text{ML}}) \sim \frac{\sqrt{B}}{\sqrt{n}}, \text{ as } n \rightarrow \infty. \quad (5.25)$$

The convergence rate in (5.25) can also be obtained from the CLT for the MLE. Comparing (5.23) and (5.25), one sees that as  $n$  gets larger, the bias of the MLE becomes negligible compared to the standard deviation. This is especially important with financial markets data, where sample sizes tend to be large.

Second, even if the MLE of a parameter  $\theta$  is unbiased, the same is not true for a nonlinear function of  $\theta$ . For example, even if  $\hat{\sigma}^2$  is unbiased for  $\sigma^2$ ,  $\hat{\sigma}$  is biased for  $\sigma$ . The reason for this is that for a nonlinear function  $g$ , in general,

$$E\{g(\hat{\theta})\} \neq g\{E(\hat{\theta})\}.$$

Therefore, it is impossible to correct for all biases.

## 5.11 Likelihood Ratio Tests

Some readers may wish to review hypothesis testing by reading Section A.18 before starting this section.

*Likelihood ratio tests*, like maximum likelihood estimation, are based upon the likelihood function. Both are convenient, all-purpose tools that are widely used in practice.

Suppose that  $\theta$  is a parameter vector and that the null hypothesis puts  $m$  equality constraints on  $\theta$ . More precisely, there are  $m$  functions  $g_1, \dots, g_m$  and the null hypothesis is that  $g_i(\theta) = 0$  for  $i = 1, \dots, m$ . It is also assumed that none of these constraints is redundant, that is, implied by the others. To illustrate redundancy, suppose that  $\theta = (\theta_1, \theta_2, \theta_3)$  and the constraints are  $\theta_1 = 0$ ,  $\theta_2 = 0$ , and  $\theta_1 + \theta_2 = 0$ . Then the constraints have a redundancy and any one of the three could be dropped. Thus,  $m = 2$ , not 3.

Of course, redundancies need not be so easy to detect. One way to check is that the  $m \times \dim(\theta)$  matrix

$$\begin{pmatrix} \nabla g_1(\theta) \\ \vdots \\ \nabla g_m(\theta) \end{pmatrix} \quad (5.26)$$

must have rank  $m$ . Here  $\nabla g_i(\theta)$  is the gradient of  $g_i$ .

As an example, one might want to test that a population mean is zero; then  $\theta = (\mu, \sigma)^T$  and  $m = 1$  since the null hypothesis puts one constraint on  $\theta$ , specifically that  $\mu = 0$ .

Let  $\hat{\theta}_{\text{ML}}$  be the maximum likelihood estimator without restrictions and let  $\hat{\theta}_{0,\text{ML}}$  be the value of  $\theta$  that maximizes  $L(\theta)$  subject to the restrictions of

the null hypothesis. If  $H_0$  is true, then  $\hat{\theta}_{0,ML}$  and  $\hat{\theta}_{ML}$  should both be close to  $\theta$  and therefore  $L(\hat{\theta}_{0,ML})$  should be similar to  $L(\hat{\theta})$ . If  $H_0$  is false, then the constraints will keep  $\hat{\theta}_{0,ML}$  far from  $\hat{\theta}_{ML}$  and so  $L(\hat{\theta}_{0,ML})$  should be noticeably *smaller* than  $L(\hat{\theta})$ .

The likelihood ratio test rejects  $H_0$  if

$$2 \left[ \log\{L(\hat{\theta}_{ML})\} - \log\{L(\hat{\theta}_{0,ML})\} \right] \geq c, \quad (5.27)$$

where  $c$  is a critical value. The left-hand side of (5.27) is twice the log of the likelihood ratio  $L(\hat{\theta}_{ML})/L(\hat{\theta}_{0,ML})$ , hence the name likelihood ratio test. Often, an *exact critical value* can be found. A critical value is exact if it gives a level that is exactly equal to  $\alpha$ . When an exact critical value is unknown, then the usual choice of the critical value is

$$c = \chi_{\alpha,m}^2, \quad (5.28)$$

where, as defined in Section A.10.1,  $\chi_{\alpha,m}^2$  is the  $\alpha$ -upper quantile value of the chi-squared distribution with  $m$  degrees of freedom.<sup>10</sup> The critical value (5.28) is only approximate and uses the fact that under the null hypothesis, as the sample size increases the distribution of twice the log-likelihood ratio converges to the chi-squared distribution with  $m$  degrees of freedom if certain assumptions hold. One of these assumptions is that the null hypothesis is *not* on the boundary of the parameter space. For example, if the null hypothesis is that a variance parameter is zero, then the null hypothesis is on the boundary of the parameter space since a variance must be zero or greater. In this case (5.27) should not be used; see Self and Liang (1987). Also, if the sample size is small, then the large-sample approximation (5.27) is suspect and should be used with caution. An alternative is to use the bootstrap to determine the rejection region. The bootstrap is discussed in Chapter 6.

Computation of likelihood ratio tests is often very simple. In some cases, the test is computed automatically by statistical software. In other cases, software will compute the log-likelihood for each model and these can be plugged into the left-hand side of (5.27).

## 5.12 AIC and BIC

An important practical problem is choosing between two or more statistical models that might be appropriate for a data set. The maximized value of the log-likelihood, denoted here by  $\log\{L(\hat{\theta}_{ML})\}$ , can be used to measure how well a model fits the data or to compare the fits of two or more models.

<sup>10</sup> The reader should now appreciate why it is essential to calculate  $m$  correctly by eliminating redundant constraints. The wrong value of  $m$  will cause an incorrect critical value to be used.



However,  $\log\{L(\hat{\boldsymbol{\theta}}_{\text{ML}})\}$  can be increased simply by adding parameters to the model. The additional parameter do not necessarily mean that the model is a better description of the data-generating mechanism, because the additional model complexity due to added parameters may simply be fitting random noise in the data, a problem that is called *overfitting*. Therefore, models should be compared both by fit to the data and by model complexity. To find a parsimonious model one needs a good tradeoff between maximizing fit and minimizing model complexity.

*AIC* (Akaike's information criterion) and *BIC* (Bayesian information criterion) are two means for achieving a good tradeoff between fit and complexity. They differ slightly and BIC seeks a somewhat simpler model than AIC. They are defined by

$$\text{AIC} = -2\log\{L(\hat{\boldsymbol{\theta}}_{\text{ML}})\} + 2p \quad (5.29)$$

$$\text{BIC} = -2\log\{L(\hat{\boldsymbol{\theta}}_{\text{ML}})\} + \log(n)p, \quad (5.30)$$

where  $p$  equals the number of parameters in the model and  $n$  is the sample size. For both criteria, “smaller is better,” since small values tend to maximize  $L(\hat{\boldsymbol{\theta}}_{\text{ML}})$  (minimize  $-\log\{L(\hat{\boldsymbol{\theta}}_{\text{ML}})\}$ ) and minimize  $p$ , which measures model complexity. The terms  $2p$  and  $\log(n)p$  are called “complexity penalties” since the penalize larger models.

The term *deviance* is often used for minus twice the log-likelihood, so  $\text{AIC} = \text{deviance} + 2p$  and  $\text{BIC} = \text{deviance} + \log(n)p$ . Deviance quantifies model fit, with smaller values implying better fit.

Generally, from a group of candidate models, one selects the model that minimizes whichever criterion, AIC or BIC, is being used. However, any model that is within 2 or 3 of the minimum value is a good candidate and might be selected instead, for example, because it is simpler or more convenient to use than the model achieving the absolute minimum. Since  $\log(n) > 2$  provided, as is typical, that  $n > 8$ , BIC penalizes model complexity more than AIC does, and for this reason BIC tends to select simpler models than AIC. However, it is common for both criteria to select the same, or nearly the same, model. Of course, if several candidate models all have the same value of  $p$ , then AIC, BIC, and  $-2\log\{L(\hat{\boldsymbol{\theta}}_{\text{ML}})\}$  are minimized by the same model.

## 5.13 Validation Data and Cross-Validation

When the same data are used both to estimate parameters and to assess fit, there is a strong tendency towards overfitting. Data contain both a *signal* and *noise*. The signal contains characteristics that are present in each sample from the population, but the noise is random and varies from sample to sample. *Overfitting* means selecting an unnecessarily complex model to fit the noise. The obvious remedy to overfitting is to diagnose model fit using data that are independent of the data used for parameter estimation. We will call the

data used for estimation the *training data* and the data used to assess fit the *validation data* or *test data*.

*Example 5.3. Estimating the expected returns of midcap stocks*

This example uses 500 daily returns on 20 midcap stocks in the `midcapD.ts` data set in R's `fEcofin` package. The data are from February 28, 1991, to December 29, 1995. Suppose we need to estimate the 20 expected returns. Consider two estimators. The first, called “separate-means,” is simply the 20 sample means. The second, “common-mean,” uses the average of the 20 sample means as the common estimator of all 20 expected returns.

The rationale behind the common-mean estimator is that midcap stocks should have similar expected returns. The common-mean estimator pools data and greatly reduces the variance of the estimator. The common-mean estimator has some bias because the true expected returns will not be identical, which is the requirement for unbiasedness of the common-mean estimator. The separate-means estimator is unbiased but at the expense of a higher variance. This is a classic example of a bias–variance tradeoff.

Which estimator achieves the best tradeoff? To address this question, the data were divided into the returns for the first 250 days (training data) and for the last 250 days (validation data). The criterion for assessing goodness-of-fit was the sum of squared errors, which is

$$\sum_{k=1}^{20} \left( \hat{\mu}_k^{\text{train}} - \bar{Y}_k^{\text{val}} \right)^2,$$

where  $\hat{\mu}_k^{\text{train}}$  is the estimator (using the training data) of the  $k$ th expected return and  $\bar{Y}_k^{\text{val}}$  is the validation data sample mean of the returns on the  $k$ th stock. The sum of squared errors are 3.262 and 0.898, respectively, for the separate-means and common-mean estimators. The conclusion, of course, is that in this example the common-mean estimator is much more accurate than using separate means.

Suppose we had used the training data also for validation? The goodness-of-fit criterion would have been

$$\sum_{k=1}^{20} \left( \hat{\mu}_k^{\text{train}} - \bar{Y}_k^{\text{train}} \right)^2,$$

where  $\bar{Y}_k^{\text{train}}$  is the training data sample mean for the  $k$ th stock and is also the separate-means estimator for that stock. What would the results have been? Trivially, the sum of squared errors for the separate-means estimator would have been 0—each mean is estimated by itself with perfect accuracy! The common-mean estimator has a sum of squared errors equal to 0.920. The inappropriate use of the training data for validation would have led to the erroneous conclusion that the separate-means estimator is more accurate.

There are compromises between the two extremes of a common mean and separate means. These compromise estimators shrink the separate means toward the common mean. Bayesian estimation, discussed in Chapter 20, is an effective method for selecting the amount of shrinkage; see Example 20.12, where this set of returns is analyzed further.  $\square$

A common criterion for judging fit is the deviance, which is  $-2$  times the log-likelihood. The deviance of the validation data is

$$-2 \log f \left( \mathbf{Y}^{\text{val}} | \hat{\boldsymbol{\theta}}^{\text{train}} \right), \quad (5.31)$$

where  $\hat{\boldsymbol{\theta}}^{\text{train}}$  is the MLE of the training data and  $\mathbf{Y}^{\text{val}}$  is the validation data.

When the sample size is small, splitting the data once into training and validation data is wasteful. A better technique is *cross-validation*, often called simply CV, where each observation gets to play both roles, training and validation.  $K$ -fold cross-validation divides the data set into  $K$  subsets of roughly equal size. Validation is done  $K$  times. In the  $k$ th validation,  $k = 1, \dots, K$ , the  $k$ th subset is the validation data and the other  $K - 1$  subsets are combined to form the training data. The  $K$  estimates of goodness-of-fit are combined, for example, by averaging them. A common choice is  $n$ -fold cross-validation, also called *leave-one-out* cross-validation. With leave-one-out cross-validation, each observation takes a turn at being the validation data set, with the other  $n - 1$  observations as the training data.

An alternative to actually using validation data is to calculate what would happen if new data could be obtained and used for validation. This is how AIC was derived. AIC is an approximation to the expected deviance of a hypothetical new sample that is independent of the actual data. More precisely, AIC approximates

$$E \left[ -2 \log f \left\{ \mathbf{Y}^{\text{new}} | \hat{\boldsymbol{\theta}}(\mathbf{Y}^{\text{obs}}) \right\} \right], \quad (5.32)$$

where  $\mathbf{Y}^{\text{obs}}$  is the observed data,  $\hat{\boldsymbol{\theta}}(\mathbf{Y}^{\text{obs}})$  is the MLE computed from  $\mathbf{Y}^{\text{obs}}$ , and  $\mathbf{Y}^{\text{new}}$  is a hypothetical new data set such that  $\mathbf{Y}^{\text{obs}}$  and  $\mathbf{Y}^{\text{new}}$  are i.i.d. Since  $\mathbf{Y}^{\text{new}}$  is not observed but has the same distribution as  $\mathbf{Y}^{\text{obs}}$ , to obtain AIC one substitutes  $\mathbf{Y}^{\text{obs}}$  for  $\mathbf{Y}^{\text{new}}$  in (5.32) and omits the expectation in (5.32). Then one calculates the effect of this substitution. The approximate effect is to reduce (5.32) by twice the number of parameters. Therefore, AIC compensates by adding  $2p$  to the deviance, so that

$$\text{AIC} = -2 \log f \left\{ \mathbf{Y}^{\text{obs}} | \hat{\boldsymbol{\theta}}(\mathbf{Y}^{\text{obs}}) \right\} + 2p, \quad (5.33)$$

which is a reexpression of (5.29).

The approximation used in AIC becomes more accurate when the sample size increases. A small-sample correction to AIC is

$$\text{AIC}_c = \text{AIC} + \frac{2p(p+1)}{n-p-1}. \quad (5.34)$$

Financial markets data sets are often large enough that the correction term  $2p(p+1)/(n-p-1)$  is small, so that AIC is adequate and  $\text{AIC}_c$  is not needed. For example, if  $n = 200$ , then  $2p(p+1)/(n-p-1)$  is 0.12, 0.21, 0.31, and 0.44 and for  $p = 3, 4, 5$ , and 6, respectively. Since a difference less than 1 in AIC values is usually considered as inconsequential, the correction would have little effect when comparing models with 3 to 6 parameters when  $n$  is at least 200. Even more dramatically, when  $n$  is 500, then the corrections for 3, 4, 5, and 6 parameters are only 0.05, 0.08, 0.12, and 0.17.

Traders usually develop trading strategies using a set of historical data and then test the strategies on new data. This is called *back-testing* and is a form of validation.

## 5.14 Fitting Distributions by Maximum Likelihood

Our first application of maximum likelihood will be to estimate parameters in univariate marginal models. Suppose that  $Y_1, \dots, Y_n$  is an i.i.d. sample from a  $t$ -distribution. Let

$$f_{t,\nu}^{\text{std}}(y \mid \mu, \sigma) \quad (5.35)$$

be the density of the standardized  $t$ -distribution with  $\nu$  degrees of freedom and with mean  $\mu$  and standard deviation  $\sigma$ . Then the parameters  $\nu$ ,  $\mu$ , and  $\sigma$  are estimated by maximizing

$$\sum_{i=1}^n \log \left\{ f_{t,\nu}^{\text{std}}(Y_i \mid \mu, \sigma) \right\} \quad (5.36)$$

using any convenient optimization software. Estimation of other models is similar.

In the following examples,  $t$ -distributions and generalized error distributions are fit.

### *Example 5.4. Fitting a $t$ -distribution to changes in risk-free returns*

This example uses one of the time series in Chapter 4, the changes in the risk-free returns that has been called `diffrf`.

First we will fit the  $t$ -distribution to the changes in the risk-free returns using R. There are two R functions that can be used for this purpose, `stdFit` and `fitdistr`. They differ in their choices of the scale parameter. `stdFit` fits the standardized  $t$ -distribution,  $t^{\text{std}}$ , and returns the estimated standard deviation, which is called “sd” (as well as the estimated mean and estimated df). `stdFit` gives the following output for the variable `diffrf`.

```
$minimum
[1] -693.2

$estimate
      mean      sd      nu
0.001214 0.072471 3.334112
```

Thus, the estimated mean is 0.001214, the estimated standard deviation is 0.07247, and the estimated value of  $\nu$  is 3.334. The function `stdFit` minimizes minus the log-likelihood and the minimum value is  $-693.2$ , or, equivalently, the maximum of the log-likelihood is 693.2.

`fitdistr` fits the classical  $t$ -distribution and returns the standard deviation times  $\sqrt{(\nu - 2)/\nu}$ , which is called `s` in the R output and is the parameter called “the scale parameter” in Section 5.5.2 and denoted there by  $\lambda$ . `fitdistr` gives the following output for `diffrrf`.

```
      m      s      df
0.001224 0.045855 3.336704
(0.002454) (0.002458) (0.500010)
```

The standard errors are in parentheses below the estimates and were computed using observed Fisher information. The estimates of the scale parameter by `stdFit` and `fitdistr` agree since  $0.045855 = \sqrt{1.3367/3.3367} \times 0.072471$ . Minor differences in the estimates of  $\mu$  and  $\nu$  are due to numerical error and are small relative to the standard errors.

AIC for the  $t$ -model is  $(2)(-693.2) + (2)(3) = -1380.4$  while BIC is  $(2)(-693.2) + \log(515)(3) = -1367.667$  because the sample size is 515.

Because the sample size is large, by the central limit theorem for the MLE, the estimates are approximately normally distributed and this can be used to construct confidence intervals. Using the estimate and standard error above, a 95% confidence interval for  $\lambda$  is

$$0.045855 \pm (1.96)(0.002458)$$

since  $z_{0.025} = 1.96$ .

□

*Example 5.5. Fitting an F-S skewed  $t$ -distribution to changes in risk-free returns*

Next the F-S skewed  $t$ -distribution is fit to `diffrrf` using the R function `sstdFit`. The results are

```
$minimum
[1] -693.2
```

```
$estimate
      mean      sd      nu      xi
0.001180 0.072459 3.335534 0.998708
```

The shape parameter  $\xi$  is nearly 1 and the maximized value of the log-likelihood is the same as for the symmetric  $t$ -distribution, which imply that a symmetric  $t$ -distribution provides as good a fit as a skewed  $t$ -distribution.  $\square$

*Example 5.6. Fitting a generalized error distribution to changes in risk-free returns*

The fit of the generalized error distribution to `diffrrf` was obtained from the R function `gedFit` and is

```
$minimum
[1] -684.8

$estimate
[1] -3.297e-07 6.891e-02 9.978e-01
```

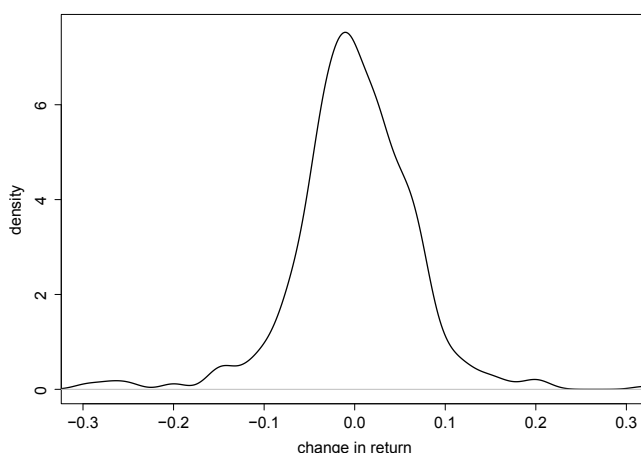
The three components of `$estimate` are the estimates of the mean, standard deviation, and  $\nu$ , respectively. The estimated shape parameter is  $\hat{\nu} = 0.998$ , which, when rounded to 1, implies a double-exponential distribution. Note that the maximum value of the likelihood is 684.8, much smaller than the value 693.2 obtained using the  $t$ -distribution. Therefore,  $t$ -distributions appear to be better models for these data compared to generalized error distributions. A possible reason for this is that, like the  $t$ -distributions, the density of the data seems to be rounded near the median; see the kernel density estimate in [Figure 5.9](#). QQ plots of `diffrrf` versus the quantiles of the fitted  $t$ - and generalized error distributions are similar, indicating that neither model has a decidedly better fit than the other. However, the QQ plot of the  $t$ -distribution is slightly more linear.

The fit to the skewed ged obtained from the R function `sgedFit` is

```
$minimum
[1] -684.8

$estimate
[1] -0.0004947 0.0687035 0.9997982 0.9949253
```

The four components of `$estimate` are the estimates of the mean, standard deviation,  $\nu$ , and  $\xi$ , respectively. These estimates again suggest that a skewed model is not needed for this example since  $\hat{\xi} = 0.995 \approx 1$ .  $\square$



**Fig. 5.9.** Kernel estimate of the probability density of `diffrf`, the changes in the risk-free returns.

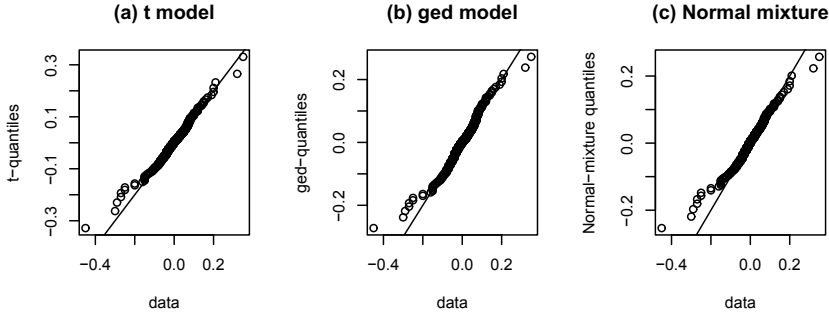
*Example 5.7. Comparing models for changes in risk-free returns*

AIC and BIC for the four models fit to the risk-free returns are reported in Table 5.1, as well as for fifth and sixth models,  $t$ -mixture and normal mixture, to be discussed next. We will ignore the mixture models for now and only consider the first four models in the table. Then, by either criterion, the  $t$ -model is best. With AIC, the skewed  $t$ -distribution is a close second, but since this model is more complex than the  $t$ -model, there is no good reason to prefer it.

**Table 5.1.** AIC and BIC for six models for the marginal distribution of `diffrf`. 1300 was added to all AIC and BIC values to improve readability.

Distribution	# Parameters	AIC	BIC
$t$	3	-80.4	-67.7
skewed $t$	4	-78.4	-61.4
ged	3	-75.6	-50.9
skewed ged	4	-61.6	-44.6
$t$ mixture	5	-82.3	-61.1
normal mixture	4	-84.2	-67.2

□



**Fig. 5.10.** (a) QQ plot of `diffrrf` versus the quantiles of a  $t_{\nu}^{\text{std}}(\mu, s^2)$  distribution with  $\mu$ ,  $s^2$ , and  $\nu$  estimated by maximum likelihood. A  $45^\circ$  line through the origin has been added for reference. (b) A similar plot for the generalized error distribution. (c) A similar plot for the normal mixture model in Example 5.8.

*Example 5.8.* Fitting a mixture model to the risk-free returns changes by maximum likelihood

The QQ plots in Figures 5.10(a) and (b) show that the risk-free returns changes have somewhat heavier tails compared to the  $t$ - and generalized error distributions.

Now consider a mixture of  $t$ -distributions as an alternative to the  $t$ - and GED models. Let  $\text{dstd}(y|\mu, s^2, \nu)$  be the value at  $y$  of the  $t_{\nu}^{\text{std}}(\mu, s^2)$  density.<sup>11</sup> Then our model for the marginal density is

$$\beta_5 \text{dstd}(y|\beta_1, \beta_2, \beta_4) + (1 - \beta_5) \text{dstd}(y|\beta_1, \beta_2 + \beta_3, \beta_4)$$

with constraints

$$\beta_2 > 0, \quad (5.37)$$

$$\beta_3 > 0, \quad (5.38)$$

$$\beta_4 > 2.1, \quad (5.39)$$

$$\beta_5 \in (0, 1). \quad (5.40)$$

Thus, the marginal density is a mixture of two  $t$ -distributions with a common mean of  $\beta_1$  and a common degrees-of-freedom parameter of  $\beta_4 = \nu$ . The first component has a variance of  $\beta_2$  and the second component has a larger variance equal to  $\beta_2 + \beta_3$ . The parameter  $\beta_5$  is the proportion of the changes in the risk-free returns coming from the first component. Since a  $t$ -distribution has an infinite variance if  $\nu \leq 2$ , the constraint  $\nu = \beta_4 > 2.1$  is imposed.

<sup>11</sup> The notation `dstd` was suggested by the name of an R function that computes this density.



A possible interpretation of this model is that the market can be in either of two possible “regimes,” the market is more volatile under the second regime than under the first,  $\beta_5$  is the probability of it being in the first regime, and  $\beta_3$  is the extra variance associated with the second regime. The somewhat outlying points in any of the three panels of Figure 5.10 would then be interpreted as data from the second regime.

AIC and BIC for this mixture model are found in the last row of Table 5.1. The mixture model has the smallest AIC among the first five models, which is evidence in its favor. However, the simpler  $t$ -model has a considerably smaller BIC value. Because of the rather small deviation from linearity in the QQ plot in Figure 5.10(a) and the large BIC value of the mixture model, our choice would be to use the simpler  $t$ -model rather than the  $t$ -mixture model.

To find the MLE for the mixture model, an R function was written to compute the log-likelihood. This function used the R function `dstd` to compute the densities of the two components. Then minus the log-likelihood was minimized using R’s minimization function `optim`, which has several different optimization algorithms—the “L-BFGS-B” algorithm was used because this algorithm allowed us to put lower and upper bounds on parameters to implement constraints (5.37)–(5.40). Optimization algorithms start at user-supplied initial values and then iteratively improve these starting values to locate a function’s minimum. The `optim` function was used 15 times starting at randomly chosen values—the starting values were uniformly distributed over ranges,  $(-0.01, 0.01)$ ,  $(0.001, 0.05)$ ,  $(0.001, 0.05)$ ,  $(2.1, 60)$ , and  $(0, 1)$  for  $\beta_1, \dots, \beta_5$ , respectively. The values of AIC, BIC, and the parameter estimates at the 15 final values are:

iter	AIC	BIC	beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[4]start
[1,]	-1382.3	-1361.1	0.0018379	0.048386	0.10908	37.218	0.88010	37.218
[2,]	-1381.5	-1360.3	0.0016881	0.051003	0.11491	10.950	0.89835	10.954
[3,]	-1382.3	-1361.1	0.0018038	0.048847	0.10994	24.791	0.88343	24.791
[4,]	-1382.2	-1361.0	0.0017831	0.049163	0.11055	20.550	0.88574	20.552
[5,]	-1382.3	-1361.1	0.0018538	0.048117	0.10873	54.093	0.87815	54.093
[6,]	-1382.3	-1361.1	0.0018257	0.048531	0.10934	32.153	0.88116	32.153
[7,]	-1382.3	-1361.1	0.0018567	0.048077	0.10868	58.141	0.87787	58.141
[8,]	-1382.3	-1361.1	0.0018414	0.048307	0.10894	40.751	0.87956	40.751
[9,]	-1382.3	-1361.1	0.0018272	0.048447	0.10919	34.958	0.88054	34.963
[10,]	-1382.3	-1361.1	0.0018421	0.048259	0.10888	42.918	0.87909	42.920
[11,]	-1382.3	-1361.1	0.0018491	0.048108	0.10868	54.481	0.87809	54.481
[12,]	-1382.3	-1361.1	0.0018303	0.048403	0.10914	36.640	0.88029	36.641
[13,]	-1382.2	-1361.0	0.0017822	0.049174	0.11056	20.399	0.88581	20.410
[14,]	-1233.4	-1212.2	0.0044000	0.045578	0.00010	13.641	0.89385	13.643
[15,]	-1382.3	-1361.1	0.0018399	0.048152	0.10877	50.468	0.87843	50.468

The last column gives the randomly chosen starting value of  $\beta_4$ . Note that only 11 of the 15 final AIC values achieve the minimum<sup>12</sup> of  $-1382.3$  though two more come close at  $-1382.2$ . The degrees-of-freedom parameter (`beta[4]`) is very poorly determined and rarely moves much from its starting value. If this

<sup>12</sup> Since the number of parameters is fixed, minimizing AIC is equivalent to maximizing the likelihood.

parameter starts at too low a value, as in cases 2 and 14, then the global minimum of AIC may not be reached. The problem is due to having three parameters,  $\beta_4 = \nu$ ,  $\beta_3$  and  $\beta_5$ , to determine tail weight, in contrast to the  $t$ -distribution with only a single tail -eight parameter  $\nu$ .

Thus, three tail-weight parameters seem to be too many. The question then is whether one tail-weight parameter (as with the simple  $t$ -model) is enough. To address this question, one can fit a two-component normal mixture model similar to the two-component  $t$ -mixture model just fit. In fact, the normal mixture model is the  $t$ -mixture model with  $\nu = \infty$ . Fixing  $\nu$  reduces the number of tail-weight parameters from three to two. The MLE was found using `optim` in R and was stable—10 random starting values all reached the same final value.<sup>13</sup>

The AIC and BIC values for the normal mixture model are in [Table 5.1](#). We see that the normal mixture model is best by AIC and second best by BIC, and for both criteria it is better than the  $t$ -mixture model. [Figure 5.10\(c\)](#) is a QQ plot for the two-component normal mixture model.<sup>14</sup> Notice that it is similar to the QQ plots for the  $t$ - and GED models shown in panels (a) and (b).

The results of this example are essentially negative. We haven't been able to improve upon the simple  $t$ -model. However, the negative results are reassuring. A good way to test whether a model fits the data adequately is to see if more complex models can achieve a better fit. If the more complex models cannot achieve substantially better fits, then this is evidence that the simpler model is adequate. Thus, there is some assurance that the simple  $t$ -model provides an adequate fit to the changes in the risk-free returns.

This example has illustrated several important concepts. The first is that maximum likelihood is a very general estimation method that is suitable for a wide variety of parametric models. The reason for this is that there are general-purpose optimization functions such as `optim` that can be used to find the MLE whenever one can write a function to compute the log-likelihood. The second is that unstable estimates whose final values depend heavily on the starting values can occur. When, as here, very different final estimates achieve nearly the same value of the log-likelihood, this is a sign of having too many parameters, a problem called *overparameterization*.

The third concept illustrated by this example is the somewhat limited practical value of asymptotic concepts such as polynomial versus exponen-

<sup>13</sup> One minor computational difficulty was that, during the iteration, the standard deviations of the components sometimes became too small and the R function `dnorm` that computes the normal density returned infinite values. This problem was solved by putting lower bounds on the standard deviations. The final estimates were above these bounds, showing that the lower bounds did not affect the final result. This problem illustrates how numerical computation of an MLE is not fool-proof and requires some care, but this is true of many numerical methods.

<sup>14</sup> The quantiles of the normal mixture model were obtained from the R function `qnorMix` in the `norMix` package.

tial tails and the index of a polynomial tail. Remember that these quantities describe tail behavior only in the limit as  $|x| \rightarrow \infty$ . It takes a long time to get to  $\infty$ ! In the range of  $x$ -values relevant in practice, a distribution with asymptotically light tails may appear heavy-tailed. The tail weight of any finite mixture model is no greater than the heaviest tail weight among its components.<sup>15</sup> Therefore, any finite normal mixture model has the very light tail of a normal distribution. Nonetheless, in this example a light-tailed normal mixture model was quite similar to a polynomial-tailed  $t_4$  distribution and to an exponentially tailed generalized error distribution. □

*Example 5.9. A-C skewed  $t$ -distribution fit to pipeline flows*

This example uses the daily flows in natural gas pipelines introduced in Example 4.3. Recall that all three distributions are left-skewed. There are many well-known parametric families of right-skewed distributions, such as, the gamma and log-normal distributions, but there are not as many families of left-skewed distributions. The F-S skewed  $t$ - and A-C skewed  $t$ -distributions, which contain both right- and left-skewed distributions, are important exceptions. In this example, the A-C skewed  $t$ -distribution will be used, though the F-S skewed  $t$ -distributions could have been used instead.

Figure 5.11 has one row of plots for each variable. The left plots have two density estimates, an estimate using the Azzalini–Capitanio skewed  $t$ -distribution (solid) and a KDE (dashed). The right plots are QQ plots using the fitted skewed  $t$ -distributions.

The flows in pipelines 1 and 2 are fit reasonably well by the A-C skewed  $t$ -distribution. This can be seen in the agreement between the parametric density estimates and the KDEs and in the nearly straight patterns in the QQ plots. The flows in pipeline 3 have a KDE with either a wide, flat mode or, perhaps, two modes. This pattern cannot be accommodated very well by the A-C skewed  $t$ -distributions. The result is less agreement between the parametric and KDE fits and a curved QQ plot. Nonetheless, a skewed  $t$ -distribution might be an adequate approximation for some purposes.

For the flows in pipeline 1, the MLEs are

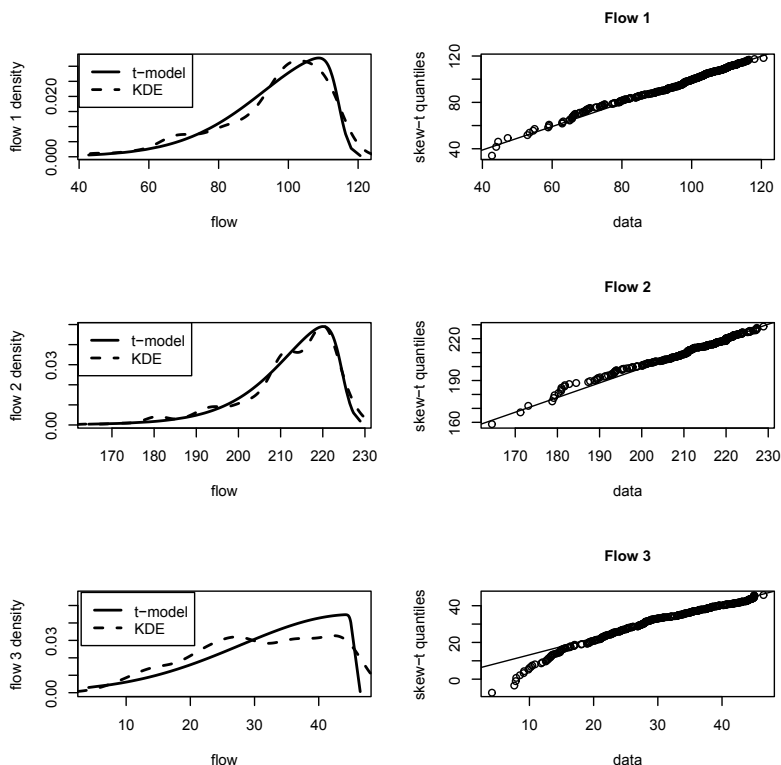
location	scale	shape	df
114.50	22.85	-9.17	15.65

and the standard errors are

location	scale	shape	df
0.637	1.849	1.977	14.863

---

<sup>15</sup> Note the assumption that there are only a finite number of components. Continuous mixtures of normal distributions include the  $t$ -distributions and other heavy-tailed distributions.



**Fig. 5.11.** Parametric (solid) and nonparametric (dashed) density estimates for daily flows in three pipelines (left) and QQ plots for the parametric fits (right). The reference lines go through the first and third quartiles.

Notice that the estimated shape parameter ( $\alpha$ ) of the A-C family is very negative, with a magnitude over four times its standard error. This is strong evidence of a highly left-skewed distribution and is in agreement with the histograms and KDEs.

For the flows in pipeline 2, the MLEs are

location	scale	shape	df
224.57	14.33	-6.43	6.58

and the standard errors are

location	scale	shape	df
0.517	1.322	1.091	2.800

Thus, in comparison with pipeline 1, pipeline 2 has higher average flows, less variability, and less skewness.

For pipeline 3, the MLEs are

location	scale	shape	df
45.5	18.1	-42.9	10228.0

The function `st.mle` in R does not return standard errors for pipeline 3 flows because of numerical problems. The difficulty may be the very large value of `df` (the MLE of the degrees-of-freedom parameter).<sup>16</sup> This value suggests that the skewed-normal distribution, which corresponds to `df` equal to  $\infty$ , should be used instead of the skewed  $t$ -distribution. For the skewed-normal fit to pipeline 3 flows, the MLEs are

location	scale	shape
45.4	17.9	-38.1

and the standard errors are

location	scale	shape
0.233	0.710	17.271

The estimates for skewed-normal fit are very close to those for skewed- $t$  fit, at least relative to the standard errors of the former.

□

## 5.15 Profile Likelihood

Profile likelihood is a technique based on the likelihood ratio test introduced in Section 5.11. Profile likelihood is used to create confidence intervals and is often a convenient way to find a maximum likelihood estimator. Suppose the parameter vector is  $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2)$ , where  $\theta_1$  is a scalar parameter and the vector  $\boldsymbol{\theta}_2$  contains the other parameters in the model. The profile log-likelihood for  $\theta_1$  is

$$L_{\max}(\theta_1) = \max_{\boldsymbol{\theta}_2} L(\theta_1, \boldsymbol{\theta}_2). \quad (5.41)$$

The right-hand side of (5.41) means the  $L(\theta_1, \boldsymbol{\theta}_2)$  is maximized over  $\boldsymbol{\theta}_2$  with  $\theta_1$  fixed to create a function of  $\theta_1$  only. Define  $\hat{\boldsymbol{\theta}}_2(\theta_1)$  as the value of  $\boldsymbol{\theta}_2$  that maximizes the right-hand side of (5.41).

The MLE of  $\theta_1$  is the value,  $\hat{\theta}_1$ , that maximizes  $L_{\max}(\theta_1)$  and the MLE of  $\boldsymbol{\theta}_2$  is  $\hat{\boldsymbol{\theta}}_2(\hat{\theta}_1)$ . Let  $\theta_{0,1}$  be a hypothesized value of  $\theta_1$ . By the theory of likelihood ratio tests in Section 5.11, one accepts the null hypothesis  $H_0 : \theta_1 = \theta_{0,1}$  if

$$L_{\max}(\theta_{0,1}) > L_{\max}(\hat{\theta}_1) - \frac{1}{2}\chi_{\alpha,1}^2. \quad (5.42)$$

<sup>16</sup> A more recent version of R does not even return an estimate when fitting the skewed  $t$ -distribution to these data with the `st.mle` function.

Here  $\chi_{\alpha,1}^2$  is the  $\alpha$ -upper quantile of the chi-squared distribution with one degree of freedom. The profile likelihood confidence interval (or, more properly, confidence region since it may not be an interval) for  $\theta_1$  is the set of all null values that would be accepted, that is,

$$\left\{ \theta_1 : L_{\max}(\theta_1) > L_{\max}(\hat{\theta}_1) - \frac{1}{2}\chi_{\alpha,1}^2 \right\}. \quad (5.43)$$

The profile likelihood can be defined for a subset of the parameters, rather than for just a single parameter, but this topic will not be pursued here.

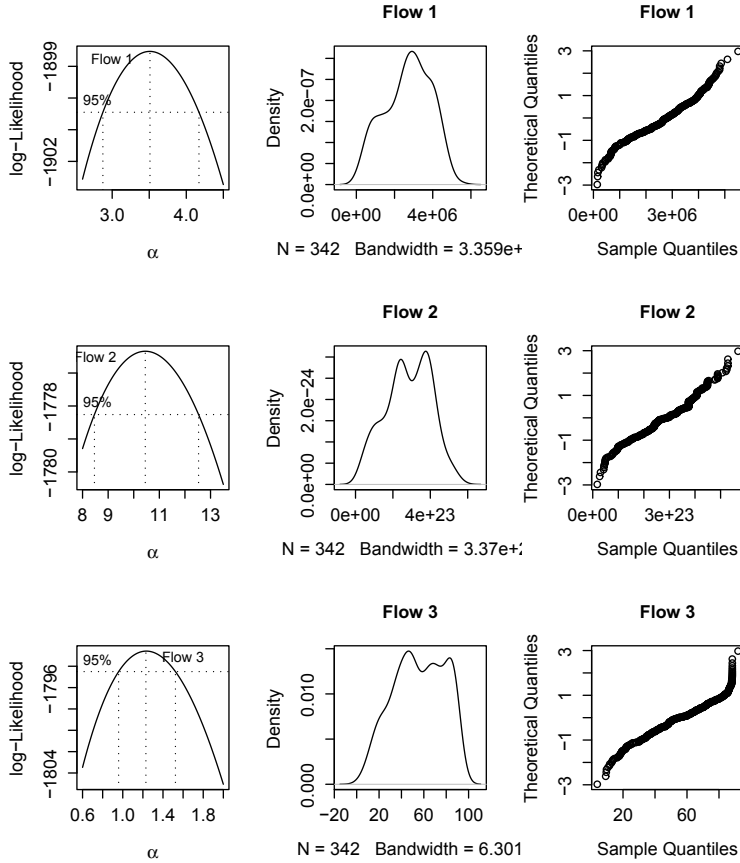
*Example 5.10. Estimating a Box–Cox transformation*

An automatic method for estimating the transformation parameter for a Box–Cox transformation assumes that for some values of  $\alpha$ ,  $\mu$ , and  $\sigma$ , the transformed data  $Y_1^{(\alpha)}, \dots, Y_n^{(\alpha)}$  are i.i.d.  $N(\mu, \sigma^2)$ -distributed. All three parameters can be estimated by maximum likelihood. For a fixed value of  $\alpha$ ,  $\hat{\mu}$  and  $\hat{\sigma}$  are the sample mean and variance of  $Y_1^{(\alpha)}, \dots, Y_n^{(\alpha)}$  and these values can be plugged into the log-likelihood to obtain the profile log-likelihood for  $\alpha$ . This can be done with the function `boxcox` in R's `MASS` package, which plots the profile log-likelihood with confidence intervals.

Estimating  $\alpha$  by the use of profile likelihood will be illustrated using the data on gas pipeline flows. [Figure 5.12](#) shows the profile log-likelihoods and the KDEs and normal QQ plots of the flows transformed using the MLE of  $\alpha$ . The KDE used `adjust` = 1.5 to smooth out local bumpiness seen with the default bandwidth. For the flows in pipeline 1, the MLE is  $\hat{\alpha} = 3.5$ . Recall that in [Example 4.3](#), we saw by trial-and-error that  $\alpha$  between 3 and 4 was best for symmetrizing the data. It is gratifying to see that maximum likelihood corroborates this choice. The QQ plots show that the Box–Cox transformed flows have light tails. Light tails are not usually considered to be a problem and are to be expected here since the pipeline flows are bounded, below by 0 and above by the capacity of the pipeline.

□

It is worth pointing out that we have now seen two distinct methods for accommodating the left skewness in the pipeline flows, modeling the untransformed data by a skewed  $t$ -distribution ([Example 5.9](#)) and Box–Cox transformation to a normal distribution ([Example 5.10](#)). A third method would be to forego parametric modeling and use the kernel density estimation. This is not an atypical situation; often data can be analyzed in several different, but equally appropriate, ways.



**Fig. 5.12.** Profile log-likelihoods and 95% confidence intervals for the parameter  $\alpha$  of the Box-Cox transformation (left), KDEs of the transformed data (middle column), and normal plots of the transformed data (right).

## 5.16 Robust Estimation

Although maximum likelihood estimators have many attractive properties, they have one serious drawback of which anyone using them should be aware. Maximum likelihood estimators can be very sensitive to the assumptions of the statistical model. For example, the MLE of the mean of a normal population is the sample mean and the MLE of  $\sigma^2$  is the sample variance, except with the minor change of a divisor of  $n$  rather than  $n-1$ . The sample mean and variance are efficient estimators when the population is truly normally distributed, but these estimators are very sensitive to outliers. Because these estimators are averages of the data and the squared deviations from the mean, respectively, a single outlier in the sample can drive the sample mean and variance to wildly

absurd values if the outlier is far enough removed from the other data. Extreme outliers are nearly impossible with exactly normally distributed data, but if the data are only approximately normal with heavier tails than the normal distribution, then outliers are more probable and, when they do occur, more likely to be extreme. Therefore, the sample mean and variance can be very inefficient estimators. Statisticians say that the MLE is not *robust* to mild deviations from the assumed model. This is bad news and has led researchers to find estimators that are robust.

A robust alternative to the sample mean is the *trimmed mean*. An  $\alpha$ -trimmed mean is computed by ordering the sample from smallest to largest, removing the fraction  $\alpha$  of the smallest and the same fraction of the largest observations, and then taking the mean of the remaining observations. The idea behind trimming is simple and should be obvious: The sample is trimmed of extreme values before the mean is calculated. There is a mathematical formulation of the  $\alpha$ -trimmed mean. Let  $k = n\alpha$  rounded<sup>17</sup> to an integer;  $k$  is the number of observations removed from both ends of the sample. Then the  $\alpha$ -trimmed mean is

$$\bar{X}_\alpha = \frac{\sum_{i=k+1}^{n-k} Y_{(i)}}{n - 2k},$$

where  $Y_{(i)}$  is the  $i$ th order statistic. Typical values of  $\alpha$  are 0.1, 0.15, 0.2, and 0.25. As  $\alpha$  approaches 0.5, the  $\alpha$ -trimmed mean approaches the sample median, which is the 0.5-sample quantile.

*Dispersion* refers to the variation in a distribution or sample. The sample standard deviation is the most common estimate of dispersion, but as stated it is nonrobust. A robust estimator of dispersion is the *MAD* (*median absolute deviation*) estimator, defined as

$$\hat{\sigma}^{\text{MAD}} = 1.4826 \times \text{median}\{|Y_i - \text{median}(Y_i)|\}. \quad (5.44)$$

This formula should be interpreted as follows. The expression “ $\text{median}(Y_i)$ ” is the sample median,  $|Y_i - \text{median}(Y_i)|$  is the absolute deviation of the observations from their median, and  $\text{median}\{|Y_i - \text{median}(Y_i)|\}$  is the median of these absolute deviations. For normally distributed data, the  $\text{median}\{|Y_i - \text{median}(Y_i)|\}$  estimates not  $\sigma$  but rather  $\Phi^{-1}(0.75)\sigma = \sigma/1.4826$ , because for normally distributed data the  $\text{median}\{|Y_i - \text{median}(Y_i)|\}$  will converge to  $\sigma/1.4826$  as the sample size increases. Thus, the factor 1.4826 in equation (5.44) calibrates  $\hat{\sigma}^{\text{MAD}}$  so that it estimates  $\sigma$  when applied to normally distributed data.

$\hat{\sigma}^{\text{MAD}}$  does not estimate  $\sigma$  for a nonnormal population. It does measure dispersion, but not dispersion as measured by the standard deviation. But this is just the point. For nonnormal populations the standard deviation is very sensitive to the tails of the distribution and does not tell us much about the dispersion in the central range of the distribution, just in the tails.

<sup>17</sup> Definitions vary and the rounding could be either upward or to the nearest integer.



In R, `mad(x)` computes (5.44). Some authors define MAD to be  $\text{median}\{|Y_i - \text{median}(Y_i)|\}$ , that is, without 1.4826. Here the notation  $\hat{\sigma}^{\text{MAD}}$  is used to emphasize the standardization by 1.4826 in order to estimate a normal standard deviation.

An alternative to using robust estimators is to assume a model where outliers are more probable. Then the MLE will automatically downweight outliers. For example, the MLE of the parameters of a  $t$ -distribution is much more robust to outliers than the MLE of the parameters of a normal distribution.

## 5.17 Transformation Kernel Density Estimation with a Parametric Transformation

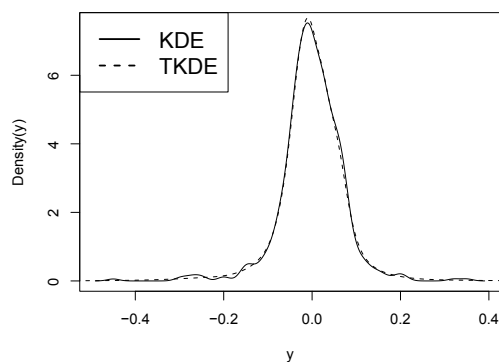
We saw in Section 4.8 that the transformation kernel density estimator (TKDE) can avoid the bumps seen when the ordinary KDE is applied to skewed data. The KDE also can exhibit bumps in the tails when both tails are long, as is common with financial markets data. An example is the variable `diffrrf` whose KDE is in Figure 5.9. For such data, the TKDE needs a transformation that is convex to the right of the mode and concave to the left of the mode. There are many such transformations, and in this section we will use some facts from probability theory, as well as maximum likelihood estimation, to select a suitable one.

The key ideas used here are that (1) normally distributed data have light tails and are suitable for estimation with the KDE, (2) it is easy to transform data to normality if one knows the CDF, and (3) the CDF can be estimated by maximum likelihood. If a random variable has a continuous distribution  $F$ , then  $F(X)$  has a uniform distribution and  $\Phi^{-1}\{F(X)\}$  has an  $N(0, 1)$  distribution; here  $\Phi$  is the standard normal CDF. Of course, in practice  $F$  is unknown, but one can estimate  $F$  parametrically, assuming, for example, that  $F$  is some  $t$ -distribution. It is not necessary that  $F$  actually be a  $t$ -distribution, only that a  $t$ -distribution can provide a reasonable enough fit to  $F$  in the tails so that an appropriate transformation is selected. If it was known that  $F$  was a  $t$ -distribution, then, of course, there would be no need to use a KDE or TKDE to estimate its density. The transformation to use in the TKDE is  $g(y) = \Phi^{-1}\{F(y)\}$ , which has inverse  $g^{-1}(x) = F^{-1}\{\Phi(x)\}$ . The derivative of  $g$  is needed to compute the TKDE and is

$$g'(y) = \frac{f(y)}{\phi[\Phi^{-1}\{F(y)\}]}.$$

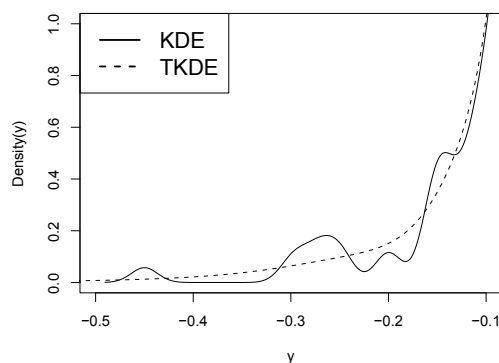
### Example 5.11. TKDE for risk-free returns

This example uses the changes in the risk-free returns in Figure 4.3. We saw in Section 5.14 that these data are reasonably well fit by a  $t$ -distribution



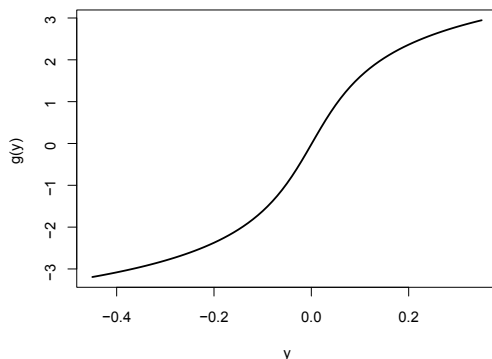
**Fig. 5.13.** Kernel density and transformation kernel density estimates of monthly changes in the risk-free returns, January 1960 to December 2002. The data are in the `Capm` series in the `Ecdat` package in R.

with mean, standard deviation, and  $\nu$  equal to 0.00121, 0.0724, and 3.33, respectively. This distribution will be used as  $F$ . Figure 5.13 compares the ordinary KDE to the TKDE for this example. Notice that the TKDE is much smoother in the tails; this can be seen better in Figure 5.14, which gives detail on the left tail.



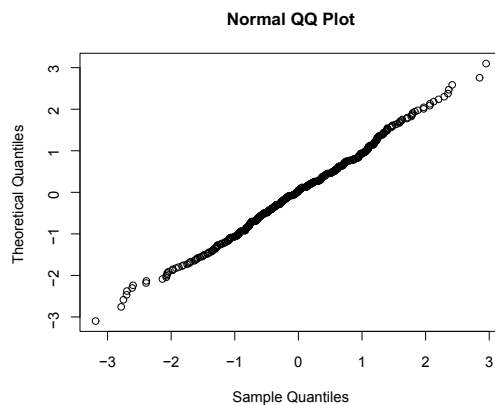
**Fig. 5.14.** Kernel density and transformation kernel density estimates of monthly changes in the risk-free returns, January 1960 to December 2002, zooming in on left tail.

The transformation used in this example is shown in [Figure 5.15](#). Notice the concave-convex shape that brings the left and right tails closer to the center and results in transformed data without the heavy tails seen in the original data. The removal of the heavy tails can be seen in [Figure 5.16](#), which is a normal plot of the transformed data.



**Fig. 5.15.** *Plot of the transformation used in Example 5.11.*

□



**Fig. 5.16.** *Normal plot of the transformed data used in Example 5.11.*

## 5.18 Bibliographic Notes

Maximum likelihood estimation and likelihood ratio tests are discussed in all textbooks on mathematical statistics, including Casella and Berger (2002) and Wasserman (2004).

Burnham and Anderson (2002) is a comprehensive introduction to model selection and is highly recommended for further reading. They also cover multimodel inference, a more advanced topic that includes *model averaging* where estimators or predictions are averaged across several models. Chapter 7 of Burnham and Anderson provides the statistical theory behind AIC as an approximate deviation of hypothetical validation data. The small-sample corrected AIC is due to Hurvich and Tsai (1989).

Buch-Larsen, Nielsen, Guillén, and Bolance (2005) and Ruppert and Wand (1992) discuss other methods for choosing the transformation when the TKDE is applied to heavy-tailed data.

The central limit theorem for the MLE is stated precisely and proved in textbooks on asymptotic theory such as Serfling (1980), van der Vaart (1998), and Lehmann (1999).

Observed and expected Fisher information are compared by Efron and Hinkley (1978), who argue that the observed Fisher information gives superior standard errors.

Box–Cox transformations were introduced by Box and Cox (1964)

## 5.19 References

- Azzalini, A., and Capitanio, A. (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistics Society, Series B*, **65**, 367–389.
- Box, G. E. P., and Dox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26** 211–246.
- Buch-Larsen, T., Nielsen, J. P., Guillén, M., and Bolance, C. (2005), Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics*, **39**, 503–518.
- Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference*, Springer, New York.
- Casella, G. and Berger, R. L. (2002) *Statistical Inference*, 2nd ed., Duxbury/Thomson Learning, Pacific Grove, CA.
- Efron, B., and Hinkley, D. V. (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**, 457–487.

- Fernandez, C., and Steel, M. F. J. (1998) On Bayesian Modelling of fat tails and skewness, *Journal of the American Statistical Association*, **93**, 359–371.
- Hurvich, C. M., and Tsai, C-L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
- Lehmann, E. L. (1999) *Elements of Large-Sample Theory*, Springer-Verlag, New York.
- Ruppert, D., and Wand, M. P. (1992) Correction for kurtosis in density estimation. *Australian Journal of Statistics*, **34**, 19–29.
- Self, S. G., and Liang, K. Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association*, **82**, 605–610.
- Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*, Wiley, New York.
- van der Vaart, A. W. (1998) *Asymptotic Statistics*, Cambridge University Press, Cambridge.
- Wasserman, L. (2004) *All of Statistics*, Springer, New York.

## 5.20 R Lab

### 5.20.1 Earnings Data

Run the following R code to find a symmetrizing transformation for 1998 earnings data from the Current Population Survey. The code looks at the untransformed data and the square-root and log-transformed data. The transformed data are compared by normal plots, boxplots, and kernel density estimates.

```
library("Ecdat")
?CPSch3
data(CPSch3)
dimnames(CPSch3)[[2]]

male.earnings = CPSch3[CPSch3[,3]=="male",2]
sqrt.male.earnings = sqrt(male.earnings)
log.male.earnings = log(male.earnings)

par(mfrow=c(2,2))
qqnorm(male.earnings,datax=T,main="untransformed")
qqnorm(sqrt.male.earnings,datax=T,main="square-root transformed")
qqnorm(log.male.earnings,datax=T,main="log-transformed")

par(mfrow=c(2,2))
boxplot(male.earnings,main="untransformed")
boxplot(sqrt.male.earnings,main="square-root transformed")
boxplot(log.male.earnings,main="log-transformed")
```

```
par(mfrow=c(2,2))
plot(density(male.earnings),main="untransformed")
plot(density(sqrt(male.earnings),main="square-root transformed")
plot(density(log(male.earnings),main="log-transformed")
```

**Problem 1** Which of the three transformation provides the most symmetric distribution? Try other powers beside the square root. Which power do you think is best for symmetrization? You may include plots with your work if you find it helpful to do that.

Next, you will estimate the Box–Cox transformation parameter by maximum likelihood. The model is that the data are  $N(\mu, \sigma^2)$ -distributed after being transformed by some  $\lambda$ . The unknown parameters are  $\lambda$ ,  $\mu$ , and  $\sigma$ .

Run the following R code to plot the profile likelihood for  $\lambda$  on the grid `seq(-2, 2, 1/10)` (this is the default and can be changed). The command `boxcox` takes an R formula as input. The left-hand side of the formula is the variable to be transformed. The right-hand side is a linear model (see Chapter 12). In this application, the model has only an intercept, which is indicated by “1.” “MASS” is an acronym for “Modern Applied Statistics with S-PLUS,” a highly-regarded textbook whose fourth edition also covers R. The MASS library accompanies this book.

```
library("MASS")
windows()
boxcox(male.earnings~1)
```

The default grid of  $\lambda$  values is large, but you can zoom in on the high-likelihood region with the following:

```
boxcox(male.earnings~1,lambda = seq(.3, .45, 1/100))
```

To find the MLE, run this R code:

```
bc = boxcox(male.earnings~1,lambda = seq(.3, .45, by=1/100),interp=F)
ind = (bc$y==max(bc$y))
ind2 = (bc$y > max(bc$y) - qchisq(.95,df=1)/2)
bc$x[ind]
bc$x[ind2]
```

- Problem 2** (a) What are `ind` and `ind2` and what purposes do they serve?  
 (b) What is the effect of `interp` on the output from `boxcox`?  
 (c) What is the MLE of  $\lambda$ ?  
 (d) What is a 95% confidence interval for  $\lambda$ ?  
 (e) Modify the code to find a 99% confidence interval for  $\lambda$ .

Rather than trying to transform the variable `male.earnings` to a Gaussian distribution, we could fit a skewed Gaussian or skewed  $t$ -distribution. R code that fits a skewed  $t$  is listed below:

```
library("fGarch")
fit = sstdFit(male.earnings,hessian=T)
```

**Problem 3** *What are the estimates of the degrees-of-freedom parameter and of  $\xi$ ?*

**Problem 4** *Produce a plot of a kernel density estimate of the pdf of `male.earnings`. Overlay a plot of the skewed  $t$ -density with MLEs of the parameters. Make sure that the two curves are clearly labeled, say with a legend, so that it is obvious which curve is which. Include your plot with your work. Compare the parametric and nonparametric estimates of the pdf. Do they seem similar? Based on the plots, do you believe that the skewed  $t$ -model provides an adequate fit to `male.earnings`?*

**Problem 5** *Fit a skewed GED model to `male.earnings` and repeat Problem 4 using the skewed GED model in place of the skewed  $t$ . Which parametric model fits the variable `male.earnings` best, skewed  $t$  or skewed GED?*

### 5.20.2 DAX Returns

This section uses log returns on the DAX index in the data set `EuStockMarkets`. Your first task is to fit the standardized  $t$ -distribution (`std`) to the log returns. This is accomplished with the following R code.

Here `loglik_std` is an R function that is defined in the code. This function returns minus the log-likelihood for the `std` model. The `std` density function is computed with the function `dstd` in the `fGarch` package. Minus the log-likelihood, which is called the objective function, is minimized by the function `optim`. The L-BFGS-B method is used because it allows us to place lower and upper bounds on the parameters. Doing this avoids the errors that would be produced if, for example, a variance parameter were negative. When `optim` is called, `start` is a vector of starting values. Use R's help to learn more about `optim`. In this example, `optim` returns an object `fit_std`. The component `fig_std$par` contains the MLEs and the component `fig_std$value` contains the minimum value of the objective function.

```
data(Garch,package="Ecdat")
library("fGarch")
data(EuStockMarkets)
Y = diff(log(EuStockMarkets[,1])) # DAX
```

```
##### std #####
loglik_std = function(x) {
  f = -sum(log(dstd(Y, x[1], x[2], x[3])))
  f}
start=c(mean(Y),sd(Y),4)
fit_std = optim(start,loglik_std,method="L-BFGS-B",
  lower=c(-.1,.001,2.1),
  upper=c(.1,1,20))
print(c("MLE =",round(fit_std$par,digits=5)))
m_logL_std = fit_std$value # minus the log-likelihood
AIC_std = 2*m_logL_std+2*length(fit_std$par)
```

**Problem 6** What are the MLEs of the mean, standard deviation, and the degrees-of-freedom parameter? What is the value of AIC?

**Problem 7** Modify the code so that the MLEs for the skewed  $t$ -distribution are found. Include your modified code with your work. What are the MLEs? Which distribution is selected by AIC, the  $t$  or the skewed  $t$ -distribution?

**Problem 8** Compute and plot the TKDE of the density of the log returns using the methodology in Sections 2.8 and 3.16 of the lecture notes. The transformation that you use should be  $g(y) = \Phi^{-1}\{F(y)\}$ , where  $F$  is the  $t$ -distribution with parameters estimated in Problem 1. Include your code and the plot with your work.

**Problem 9** Plot the KDE, TKDE, and parametric estimator of the log-return density, all on the same graph. Zoom in on the right tail, specifically the region  $0.035 < y < 0.06$ . Compare the three densities for smoothness. Are the TKDE and parametric estimates similar? Include the plot with your work.

## 5.21 Exercises

1. Load the CRSPday data set in the Ecdat package and get the variable names with the commands

```
library(Ecdat)
data(CRSPday)
dimnames(CRSPday)[[2]]
```

Plot the IBM returns with the commands

```
r = CRSPday[,5]
plot(r)
```



Learn the mode and class of the IBM returns with

```
mode(r)
class(r)
```

You will see that the class of the variable `r` is “`ts`,” which means “time series.” Data of class `ts` are plotted differently than data not of this class. To appreciate this fact, use the following commands to convert the IBM returns to class `numeric` before plotting them:

```
r2 = as.numeric(r)
class (r2)
plot(r2)
```

The variable `r2` contains the same data as the variable `r`, but `r2` has class `numeric`.

Find the covariance matrix, correlation matrix, and means of GE, IBM, and Mobil with the commands

```
cov(CRSPday[,4:6])
cor(CRSPday[,4:6])
apply(CRSPday[,4:6], 2, mean)
```

Use your R output to answer the following questions:

- (a) What is the mean of the Mobil returns?
  - (b) What is the variance of the GE returns?
  - (c) What is the covariance between the GE and Mobil returns?
  - (d) What is the correlation between the GE and Mobil returns?
2. Suppose that  $Y_1, \dots, Y_n$  are i.i.d.  $N(\mu, \sigma^2)$ , where  $\mu$  is *known*. Show that the MLE of  $\sigma^2$  is

$$n^{-1} \sum_{i=1}^n (Y_i - \mu)^2.$$

3. Show that  $f^*(y|\xi)$  given by equation (5.15) integrates to  $(\xi + \xi^{-1})/2$ .
4. Let  $X$  be a random variable with mean  $\mu$  and standard deviation  $\sigma$ .
  - (a) Show that the kurtosis of  $X$  is equal to 1 plus the variance of  $\{(X - \mu)/\sigma\}^2$ .
  - (b) Show that the kurtosis of any random variable is at least 1.
  - (c) Show that a random variable  $X$  has a kurtosis equal to 1 if and only if  $P(X = a) = P(X = b) = 1/2$  for some  $a \neq b$ .
5. (a) What is the kurtosis of a normal mixture distribution that is 95%  $N(0, 1)$  and 5%  $N(0, 10)$ ?
  - (b) Find a formula for the kurtosis of a normal mixture distribution that is  $100p\%$   $N(0, 1)$  and  $100(1 - p)\%$   $N(0, \sigma^2)$ , where  $p$  and  $\sigma$  are parameters. Your formula should give the kurtosis as a function of  $p$  and  $\sigma$ .

- (c) Show that the kurtosis of the normal mixtures in part (b) can be made arbitrarily large by choosing  $p$  and  $\sigma$  appropriately. Find values of  $p$  and  $\sigma$  so that the kurtosis is 10,000 or larger.
  - (d) Let  $M > 0$  be arbitrarily large. Show that for any  $p_0 < 1$ , no matter how close to 1, there is a  $p > p_0$  and a  $\sigma$ , such that the normal mixture with these values of  $p$  and  $\sigma$  has a kurtosis at least  $M$ . This shows that there is a normal mixture arbitrarily close to a normal distribution but with a kurtosis above any arbitrarily large value of  $M$ .
6. Fit the F-N skewed  $t$ -distribution to the gas flow data. The data set is in the file `GasFlowData.csv`, which can be found on the book's website. The F-N skewed  $t$ -distribution can be fit using the function `sstdFit` in R's `fGarch` package.
  7. Suppose that  $X_1, \dots, X_n$  are i.i.d.  $\text{exponential}(\theta)$ . Show that the MLE of  $\theta$  is  $\bar{X}$ .
  8. The number of small businesses in a certain region defaulting on loans was observed for each month over a 4-year period. In the R program below, the variable `y` is the number of defaults in a month and `x` is the value for that month of an economic variable thought to affect the default rate. The function `dpois` computes the Poisson density.

```
start =c(1,1)
loglik = function(theta) {-sum(log(dpois(y,lambda=theta[1]+
  theta[2]*x)))}
mle= optim(start,loglik,hessian=T)
invFishInfo = solve(mle$hessian)
options(digits=4)
mle$par
mle$value
mle$convergence
sqrt(diag(invFishInfo))
```

The output is

```
> mle$par
[1] 28.0834 0.6884
> mle$value
[1] 150.9
> mle$convergence
[1] 0
> sqrt(diag(invFishInfo))
[1] 1.8098 0.1638
```

- (a) Describe the statistical model being used here.
- (b) What are the parameter estimates?
- (c) Find 95% confidence intervals for the parameters in the model. Use a normal approximation.

9. In this problem you will fit a  $t$ -distribution by maximum likelihood to the daily log returns for BMW. The data are in the data set `bmw` that is part of the `evir` package. Run the following code:

```
library(evir)
library(fGarch)
data(bmw)
start_bmw = c(mean(bmw),sd(bmw),4)
loglik_bmw = function(theta)
{
  -sum(log(dstd(bmw,mean=theta[1],sd=theta[2],nu=theta[3])))
}

mle_bmw = optim(start_bmw, loglik_bmw, hessian=T)
FishInfo_bmw = solve(mle_bmw$hessian)
```

Note: The R code defines a function `loglik_bmw` that is minus the log-likelihood. See Chapter 10 of *An Introduction to R* for more information about functions in R. Also, see page 59 of this manual for more about maximum likelihood estimation in R. `optim` minimizes this objective function and returns the MLE (which is `mle_bmw$par`) and other information, including the Hessian of the objective function evaluated at the MLE (because `hessian=T`—the default is not to return the Hessian).

- (a) What does the function `dstd`, and what package is it in?
  - (b) What does the function `solve` do?
  - (c) What is the estimate of  $\nu$ , the degrees-of-freedom parameter?
  - (d) What is the standard error of  $\nu$ ?
10. In this problem, you will fit a  $t$ -distribution to daily log returns of Siemens. You will estimate the degrees-of-freedom parameter graphically and then by maximum likelihood. Run the following code, which produces a  $3 \times 2$  matrix of probability plots. If you wish, add reference lines as done in Section 4.11.1.

```
data(siemens)
n=length(siemens)
par(mfrow=c(3,2))
qqplot(siemens,qt(((1:n)-.5)/n,2),ylab="t(2) quantiles",
       xlab="data quantiles")
qqplot(siemens,qt(((1:n)-.5)/n,3),ylab="t(3) quantiles",
       xlab="data quantiles")
qqplot(siemens,qt(((1:n)-.5)/n,4),ylab="t(4) quantiles",
       xlab="data quantiles")
qqplot(siemens,qt(((1:n)-.5)/n,5),ylab="t(5) quantiles",
       xlab="data quantiles")
qqplot(siemens,qt(((1:n)-.5)/n,8),ylab="t(8) quantiles",
       xlab="data quantiles")
```

```
qqplot(siemens,qt(((1:n)-.5)/n,12),ylab="t(12) quantiles",  
       xlab="data quantiles")
```

R has excellent graphics capabilities—see Chapter 12 of *An Introduction to R* for more about R graphics and, in particular, pages 67 and 72 for more information about `par` and `mfrow`, respectively.

- (a) Do the returns have lighter or heavier tails than a  $t$ -distribution with 2 degrees of freedom?
- (b) Based on the QQ plots, what seems like a reasonable estimate of  $\nu$ ?
- (c) What is the MLE of  $\nu$  for the Siemens log returns?