

# Chapter 1

## Modelling the Distribution of Stock Returns

It is difficult to model returns on assets in financial markets. The financial markets is a strongly fluctuating system that is constantly driven by external information. The feedback between the markets and the outside world and the complex interactions between traders and assets are challenging to be understood. With the ever-changing computer technology and extensive application of statistical theory, quantitative modelling for asset returns attracts the attraction of many researchers.

There is a recent interest in modelling the stock returns (i.e. the relative price changes) by elliptical models instead of Gaussian ones. For instance because the elliptically contoured distribution can capture fat tail behavior ([9, 10]). The main properties of elliptically contoured distribution can be found in [1, 2]; Their application in mean variance portfolio optimization problem is discussed in [3, 4, 5, 7]; CAPM and a Black-Scholes type option pricing formula have been derived by [8].

Most previous work on non-Gaussian distributions focuses on the tails and uses in particular the Student t distribution because of its power law tails. However, the Student t distribution is not better than the Gaussian in the central region where most returns lie. In this paper, we look for a model that fits the empirical data reasonably well both in the tails and the central region.

Suppose we have  $p$  stocks<sup>1</sup>. We denote the price of the  $j$ th stock at time  $t$  as  $S_j(t)$  for every stock  $j = 1, 2, \dots, p$ . The stock price return in the interval  $[t, t + \Delta t]$  is denoted by

$$R = (R_1, R_2, \dots, R_p)^T, \quad R_j = \frac{S_j(t + \Delta t) - S_j(t)}{S_j(t)}.$$

We will model the distribution of  $R$  using a family of elliptically contoured distribution. As will be explained shortly, the random variable  $R$  may be represented as the sum of the constant mean  $\mu$  and the product of a general non-negative scalar random variable  $\beta$  and a multivariate normal  $\varepsilon$

$$R = \mu + \beta \cdot \varepsilon, \quad \varepsilon \sim N(0, \Omega). \quad (1.1)$$

Here, the multivariate normal  $\varepsilon$  has mean zero and covariance  $\Omega$  and the non-negative scalar random variable  $\beta$  is independent of  $\varepsilon$ .

This elliptical model separates the problem of correlation (handled by the Gaussian part) from the problem of return distribution and tails (handled by the other random factor,  $\beta$ ). We use  $\varepsilon$  to get the variances and covariances right and use the scalar  $\beta$  to adjust the shape of the distribution. In order

---

<sup>1</sup>This notation comes from El-karoui [3], where  $p$  is the number of assets and  $n$  is the number of observations.

to make the covariance of  $R$  equal to the covariance of  $\varepsilon$ , we need  $E[\beta^2] = 1$ , which always is imposed. We note that this is not the standard definition of elliptical models, which generally replaces  $\varepsilon$  with a vector uniformly distributed on an ellipsoid in  $R^p$  ([1, 2, 3]). In Appendix A, we explain how we derive the family (1.1) from the standard definition of elliptical models. The derivation is based on Chu's work [11].

Below we find a formula for the probability density function (PDF) of the stock returns  $R$ , not only because it is the most direct tool for comparing the models and the data, but because it plays an important role in our Monte Carlo method for Bayesian asset allocation discussed later in Chapter 3. First, the conditional distribution for the stock returns  $R$  given  $\beta$  is the multivariate Gaussian with mean  $\mu$  and covariance matrix  $\beta^2\Omega$ . That is  $R \mid \beta \sim N(\mu, \beta^2\Omega)$ . The formula is

$$R \mid \beta \propto |\beta^2\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2}(R-\mu)^T(\beta^2\Omega)^{-1}(R-\mu)}$$

where  $|\cdot|$  is the determinant,  $R \propto h(R)$  means that  $h(R)$  is the probability density function of  $R$  up to an overall factor. Then, we let  $p(\beta)$  be the PDF of  $\beta$ . The PDF of stock return  $R$  is

$$f(R) \propto \int_{\beta \geq 0} |\beta^2\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2}(R-\mu)^T(\beta^2\Omega)^{-1}(R-\mu)} p(\beta) d\beta. \quad (1.2)$$

In Section 1.1 we discuss various specific versions of this general model. The case of constant  $\beta = 1$  is the Gaussian return. If  $\beta^2$  has the inverse Gamma distribution then  $R = \mu + \beta\varepsilon$  has the Student t distribution, which often is preferred to the Gaussian because of its fat tails. The case where  $\beta$

has a Gamma distribution seems not to be studied in the literature. But it fits actual returns better than the Gaussian or Student t models above. This is discussed in Section 1.2.

## 1.1 Three Simple Sub-families and their PDFs

### 1.1.1 The Gaussian Model

The simplest example is when  $\beta$  is a constant,  $\beta \equiv 1$ . The stock returns  $R$  have Gaussian distribution with mean  $\mu$  and covariance matrix  $\Omega$  which has the following PDF

$$p(R) \propto |\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2}(R-\mu)^T (\Omega)^{-1} (R-\mu)}. \quad (1.3)$$

### 1.1.2 The Student t Model

A multivariate Student t distribution with mean  $\mu$ , scale matrix  $\Sigma$  and  $m$  degrees of freedom has the form

$$\mu + \epsilon \sqrt{\frac{m}{\chi_m^2}}$$

where  $\epsilon$  is normal with mean zero and covariance  $\Sigma$  and  $\chi_m^2$  is Chi Square distribution with  $m$  degrees of freedom. For  $m$  integer,  $\chi_m^2$  has the distribution of  $Z_1^2 + \dots + Z_m^2$ , where the  $Z_k$  are independent standard normals.

Let  $L^2 = Z_1^2 + \dots + Z_m^2$ . Then the PDF of  $L$  is (since  $L^2 = |Z|^2$  with  $Z \in \mathbb{R}^m$ )

$$f(l) = C l^{m-1} e^{-l^2/2}. \quad (1.4)$$

The constant is determined by the requirement that  $\int_0^\infty f(l)dl = 1$ . That is

$$\frac{1}{C} = \int_0^\infty l^{m-1} e^{-l^2/2} dl . \quad (1.5)$$

We want to take  $\beta = \text{Const } L^{-1}$ . As explained above, we want to choose the constant so that  $E[\beta^2] = 1$ . For that reason, we calculate

$$E[L^{-2}] = C \int_0^\infty l^{-2} l^{m-1} e^{-l^2/2} dl = C \int_0^\infty l^{m-3} e^{-l^2/2} dl .$$

We use an integration by parts trick that is common in this setting:

$$\begin{aligned} \int_0^\infty l^{m-1} e^{-l^2/2} dl &= \int_0^\infty l^{m-2} e^{-l^2/2} l dl \\ &= - \int_0^\infty l^{m-2} \partial_l \left( e^{-l^2/2} \right) dl \\ &= (m-2) \int_0^\infty l^{m-3} e^{-l^2/2} l dl . \end{aligned}$$

And this implies that

$$E[L^{-2}] = \frac{\int_0^\infty l^{m-3} e^{-l^2/2} dl}{\int_0^\infty l^{m-1} e^{-l^2/2} dl} = \frac{1}{m-2} . \quad (1.6)$$

This implies that

$$\beta = \frac{\sqrt{m-2}}{L} \quad (1.7)$$

has  $E[\beta^2] = 1$  as required. Note that the definition of  $L$  using (1.4) and the normalization (1.7) works for any real  $m > 2$ , not only integer values.

With this we calculate the PDF of  $R = \mu + \beta\varepsilon$  where

$$\varepsilon \propto |\Omega|^{-\frac{1}{2}} e^{-\varepsilon^t \Omega^{-1} \varepsilon / 2} .$$

As explained above, we first write the PDF of  $R - \mu$  conditional on a fixed value of  $\beta$  or  $L$ , then integrate with respect to  $f(l)$  to get the overall PDF. For fixed  $L$ ,  $R - \mu = \frac{\sqrt{m-2}}{L}\varepsilon$  is normal with mean zero and covariance  $\frac{m-2}{L^2}\Omega$ . Therefore, the conditional density is (omitting constant factors involving  $p$  and  $m$ )

$$|\Omega|^{-\frac{1}{2}} l^p \exp\left(\frac{-l^2}{2(m-2)} r^t \Omega^{-1} r\right).$$

Therefore,

$$\begin{aligned} R - \mu &\propto |\Omega|^{-\frac{1}{2}} \int_0^\infty l^{p+m-1} \exp\left(\frac{-l^2}{2} \left[\frac{1}{m-2} r^t \Omega^{-1} r + 1\right]\right) dl \\ &\propto |\Omega|^{-\frac{1}{2}} \int_0^\infty \eta^{\frac{p+m}{2}-1} \exp\left(-\eta \left[\frac{1}{m-2} r^t \Omega^{-1} r + 1\right]\right) d\eta \\ &\propto |\Omega|^{-\frac{1}{2}} \left(1 + \frac{1}{m-2} r^t \Omega^{-1} r\right)^{-(p+m)/2}. \end{aligned}$$

Experts will recognize this as an instance of the multivariate Student t distribution with  $m$  degrees of freedom and scale matrix  $\Sigma = \frac{m-2}{m}\Omega$ .

### 1.1.3 The New Model

The Student t model above is a popular model for asset returns because it makes large returns (both positive and negative) much more likely than the basic Gaussian model. However, as we show in Section 1.2, this model is actually worse than the Gaussian model in the central region. Data from US equity markets consistently show return PDF's that are more sharply peaked in the center (small returns). It turns out that models of the form  $R - \mu \sim \beta\varepsilon$ , with the appropriate choice of  $\beta$  are able to fit this rather well.

If one decides that the Gaussian model under-predicts the frequency of

small returns, an obvious solution is to choose a  $\beta$  distribution that makes small  $\beta$  reasonably likely. The Student  $t$  is not good for this (we see in retrospect) because small  $\beta$  corresponds to large  $L = \sum Z_k^2$ , which is exponentially unlikely (1.4). Instead one can try a simple distribution such as the exponential  $\beta \sim f(\beta) = e^{-\beta}$ . This turns out (see below) that even in the single asset case the PDF of  $R$  becomes infinite as  $r \rightarrow 0$ , in clear contradiction to market data. This happens whenever  $f(\beta)$  approaches a non-zero value as  $\beta \rightarrow 0$ . Making  $f(\beta)$  vanish as  $\beta \rightarrow 0$  fixes that problem. This leads us to consider the distribution  $f(\beta) \propto \beta e^{-\beta}$ .

More generally, we consider  $\beta$  to come from a Gamma distribution with PDF  $f(\beta) \propto \beta^{k-1} e^{-\beta/\theta}$ . For integer  $k$ , this is the distribution of  $k$  independent exponential random variables. We will see that  $k$  between 2 and 3 fit both single and multi asset data reasonably well.

We compute the relationship between  $\theta$  and  $k$  that implies  $E[\beta^2] = 1$ . The PDF of  $\beta$  is  $\frac{1}{C} \beta^{k-1} e^{-\beta/\theta}$ , where  $C = \int_0^\infty \beta^{k-1} e^{-\beta/\theta} d\beta$ . Then

$$\begin{aligned}
E[\beta^2] &= \frac{1}{C} \int_0^\infty \beta^{k+1} e^{-\beta/\theta} d\beta \\
&= \frac{-\theta}{C} \int_0^\infty \beta^{k+1} \partial_\beta (e^{-\beta/\theta}) d\beta \\
&= \frac{\theta(k+1)}{C} \int_0^\infty \beta^k e^{-\beta/\theta} d\beta \\
&= \frac{\theta^2(k+1)k}{C} \int_0^\infty \beta^{k-1} e^{-\beta/\theta} d\beta \\
&= \theta^2 (k^2 + k) .
\end{aligned}$$

The condition then is  $\theta^2 (k^2 + k) = 1$ , which gives  $\theta = 1/\sqrt{k^2 + k}$ . Therefore,

we take  $\beta$  to have PDF

$$f(\beta) \propto \beta^{k-1} e^{-\beta\sqrt{k^2+k}}, \quad (1.8)$$

where  $k \geq 0$  is a free parameter. With this, we find the PDF for  $R - \mu$  as before. For fixed  $\beta$ ,

$$R - \mu = \beta\varepsilon \propto \mathcal{N}(0, \beta^2\Omega) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\beta^2\Omega|^{1/2}} e^{-r^t(\beta^2\Omega)^{-1}r/2}.$$

The overall density comes from this by integration over  $\beta$  using (1.8) and  $|\beta^2\Omega| = \beta^{2p} |\Omega|$  (still omitting factors depending on the constants  $p$ , and  $k$ ):

$$R - \mu \propto \int_0^\infty \frac{\beta^{k-p-1}}{|\Omega|^{1/2}} \exp\left(-\left[\beta^{-2}r^t\Omega^{-1}r/2 + \beta\sqrt{k^2+k}\right]\right) d\beta. \quad (1.9)$$

We are unable to integrate this in closed form, but is the basis for our numerical evaluations and Monte Carlo sampling of the  $R$  distribution. It would have been simpler here to omit the determinant factor  $|\Omega|$ , but we will need it later when we treat  $\Omega$  also as unknown.

The model  $R - \mu = \beta\varepsilon$  with  $\beta$  taken from the specific Gamma distribution (1.8) has the feature that the probability density (1.9) may be singular or become infinite at the origin. For convenience in this discussion we temporarily set  $\mu = 0$  and write the probability density of  $R$  as

$$f(r) = \frac{1}{C} \int_0^\infty \beta^{k-p-1} \exp\left(-\left[\beta^{-2}r^t\Omega^{-1}r/2 + \beta\sqrt{k^2+k}\right]\right) d\beta. \quad (1.10)$$

There are two main cases. If  $k > p$ , so  $k - p - 1 > -1$ , the integral converges uniformly as  $r \rightarrow 0$  and  $f(r)$  is continuous there.



The more interesting case for us is  $k < p$ . In this case the small  $r$  behavior of  $f$  may be found by re-scaling to normalize the exponent. Define  $\rho = (r^t \Omega^{-1} r)^{1/2}$  and  $b = \beta/\rho$ , so  $\beta = \rho b$ . Then

$$f(r) = \frac{\rho^{k-p}}{C} \int_0^\infty \left( b^{k-p-1} e^{-\frac{1}{2b^2}} \right) \exp \left( - \left[ b \rho \sqrt{k^2 + k} \right] \right) db .$$

Since

$$\int_0^\infty \left( b^{k-p-1} e^{-\frac{1}{2b^2}} \right) db < \infty$$

when  $k < p$ , the dominated convergence theorem implies that

$$f(r) \sim C (r^t \Omega^{-1} r)^{(k-p)/2} = O \left( |r|^{k-p} \right) \quad \text{as } r \rightarrow 0. \quad (1.11)$$

Here,  $f(r) \sim g(r)$  as  $r \rightarrow 0$  means that  $\frac{f(r)}{g(r)} \rightarrow 1$  as  $r \rightarrow 0$ .

In statistical tests below we use the above results in the following slightly weaker form

**Theorem 1.** *If  $R$  has the distribution (1.9), then if  $k > p$ ,*

$$Pr(|R - \mu| \leq r) = O(r^p) \quad \text{as } r \rightarrow 0. \quad (1.12)$$

*If  $k < p$  then*

$$Pr(|R - \mu| \leq r) = O(r^k) \quad \text{as } r \rightarrow 0. \quad (1.13)$$

**Remark 1.** The borderline case  $k = p$  has a logarithmic singularity, as is shown by a more careful argument.

**Remark 2.** The formula (1.13), with  $k < p$ , does not imply the precise statement (1.11), but it does imply that  $f(r)$  is not bounded as  $r \rightarrow 0$ . The next section presents strong empirical evidence that multi-asset stock returns

satisfy (1.13) with  $k < p$ . This implies that multi-asset return probabilities are singular at the origin.

## 1.2 Empirical distribution of the stock returns

It is well known that quite far from the simple assumption of textbook mathematical finance, the distributions of the returns of any kind of traded financial instrument (stocks, currencies, interest rates, commodities, etc.) have much fatter tails than Gaussian. [6, 5] suggest that the empirical PDF of returns on shortish time scales (say between a few minutes and a few days) can be reasonably well fitted by a Student-t distribution with the degree of freedom in the range 3 to 5. However, interestingly, our new model introduced in section 1.1.3 gives a better fit of the empirical stock returns data in the central region within  $2\sigma$ , while the exponential tail of this model is also a plausible fit to the data in the tails.

### 1.2.1 Empirical distribution of single asset

First we look at the centered and normalized relative daily returns of *SPY*, the ETF of the *S&P500*, from Jan 1993 to May 2011. The *S&P500* is defined as the sum of the market capitalizations (stock price multiplied by the number of outstanding shares) of 500 companies representative of the U.S. economy. The daily prices of *SPY* are adjusted close prices from Yahoo! Finance which are adjusted for all splits and dividends. The returns have sample mean  $2.94bps$  and sample standard deviation  $108bps$ . The centered and normalized returns have mean zero and variance one. In Figure 1.1-1.2, we plot the PDF and the cumulative distribution function (CDF) of the

centered and normalized daily returns of *SPY* comparing to the Gaussian distribution, the Student t distribution (the degree of freedom  $m = 4$ ) and the new model ( $k = 2$ ) defined in section 1.1.3 with mean zero and variance one. We observe that the new model gives the best fit within  $2\sigma$   $([-2, 2])$ , the Student t and the new model are both plausible fit to the data in the tails.

Some people may argue that the distribution of asset returns would be mostly asymmetric [12], but we find our symmetric model has its uses in many cases. For example, in Figure 1.3, we test the upside and downside tails of the centered and normalized daily returns of *SPY* , by comparing

$$P\left(\frac{R - \mu}{\sigma} > M\right)$$

and

$$P\left(\frac{R - \mu}{\sigma} < -M\right)$$

where  $M$  is the number of standard deviations away from the center. We find that by this measure the one day return distribution is reasonably symmetric.

Some people may also ask if the autocorrelation of returns for distinct periods is significant. In Figure 1.4, we show the autocorrelation function (ACF) of the daily returns of *SPY*. ACF is defined as

$$\rho(\tau) = \frac{E(R_t - \mu)(R_{t+\tau} - \mu)}{\sigma^2}$$

for any stationary process  $R_t$ . We find that returns for distinct periods (days in this case) have reasonably small correlation (negative in Lag 1). In Chapter 4 we will introduce a new stochastic volatility model in which excess returns

are not independent and is consistent with our  $\beta \cdot \varepsilon$  model with  $\beta$  having a Gamma distribution.

Next we do similar analyses for the stocks in the *S&P500* using two-minute as well as daily returns. The two-minute returns are computed from mid prices of TAQ data from Jun 20th 2007 to Sep 20th 2007 (64 business days). The actual number of stocks we studied is 486, since the data files for some of the stocks are not complete in our database. The main features above are present here as well. The new model with  $k = 3$  (instead of 2) fits return data better than Gaussian or Student t models with the degree of freedom  $m = 4$ . For example, in Figure 1.5-1.7, we plot the PDF, CDF and ACF of the centered and normalized two-minute returns of IBM. The two-minute returns of IBM have sample mean  $0.12bps$  and sample standard deviation  $9.2bps$ . The centered and normalized returns have mean zero and variance one. In Figure 1.8-1.9, we cluster the centered and normalized two-minute returns of 486 stocks in the *S&P500* into one set and plot the PDF and CDF of this set.

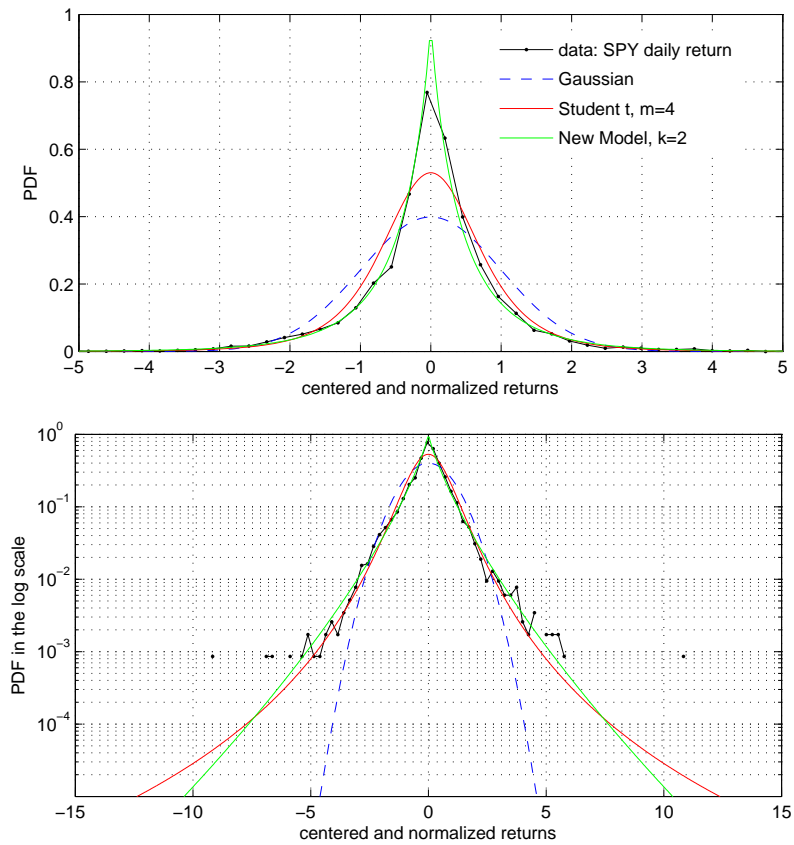


Figure 1.1: PDF of centered and normalized SPY daily returns: the data set is from Jan 1993 to May 2011.

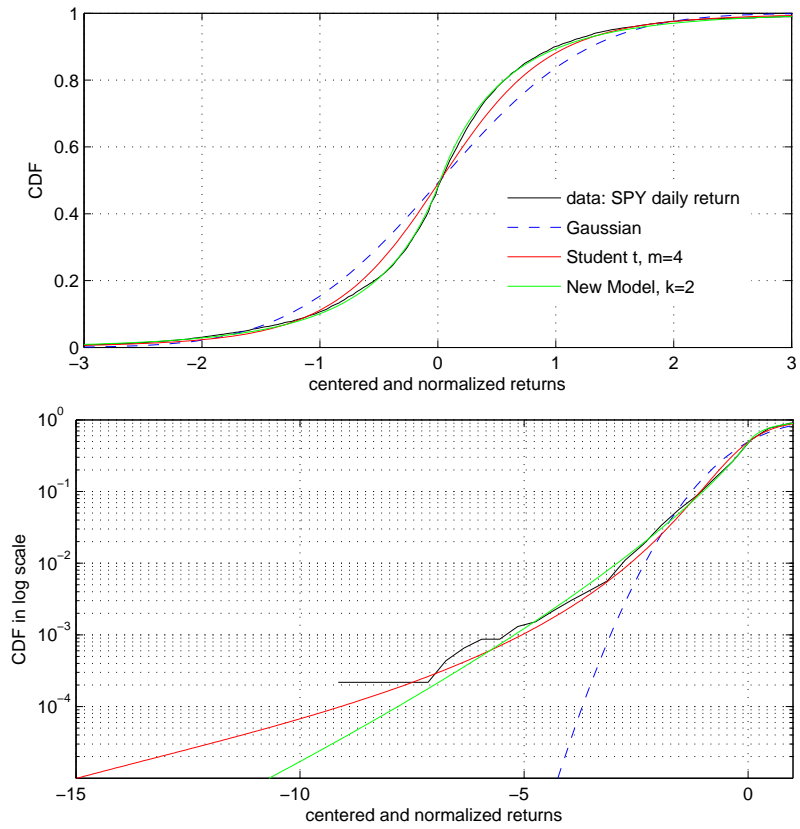


Figure 1.2: CDF of centered and normalized SPY daily returns: the data set is from Jan 1993 to May 2011.

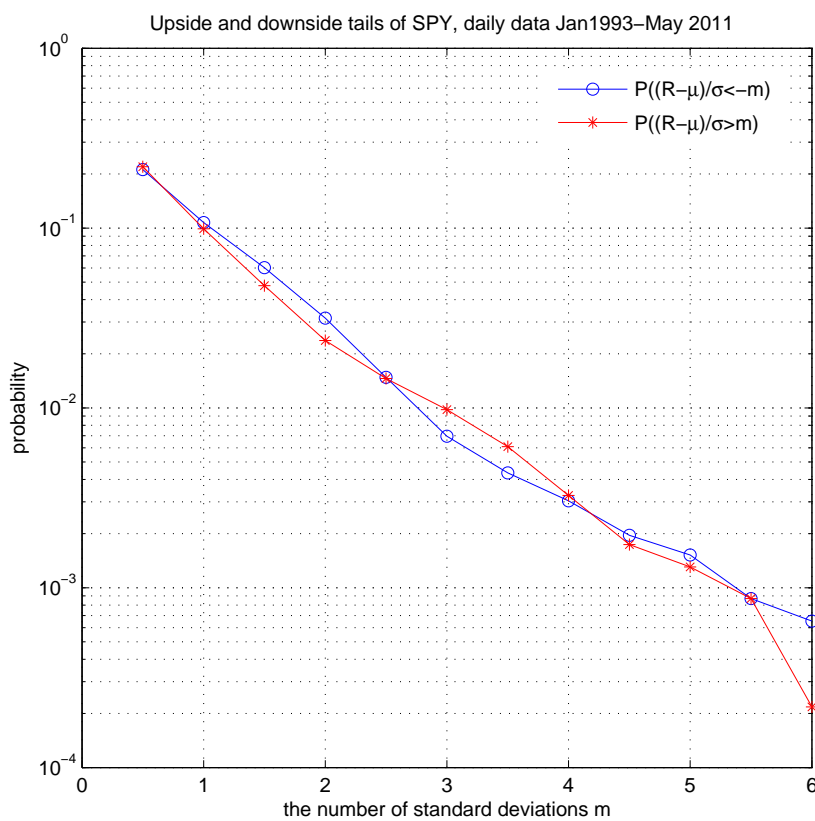


Figure 1.3: Test the upside and downside tails of the centered and normalized SPY daily returns: the data set is from Jan 1993 to May 2011.

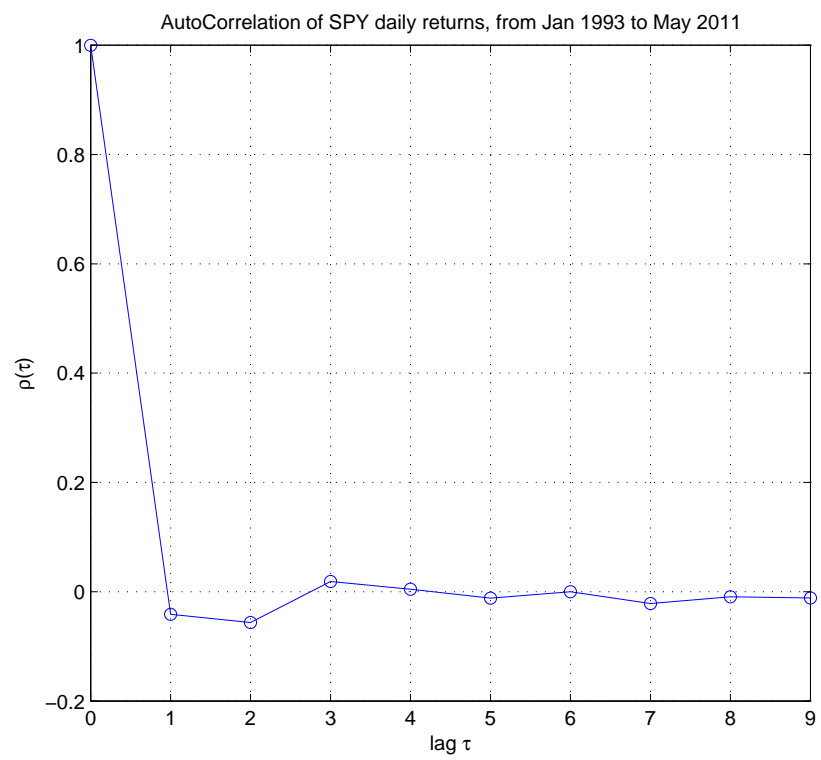


Figure 1.4: Autocorrelation of SPY daily returns: the data set is from Jan 1993 to May 2011.



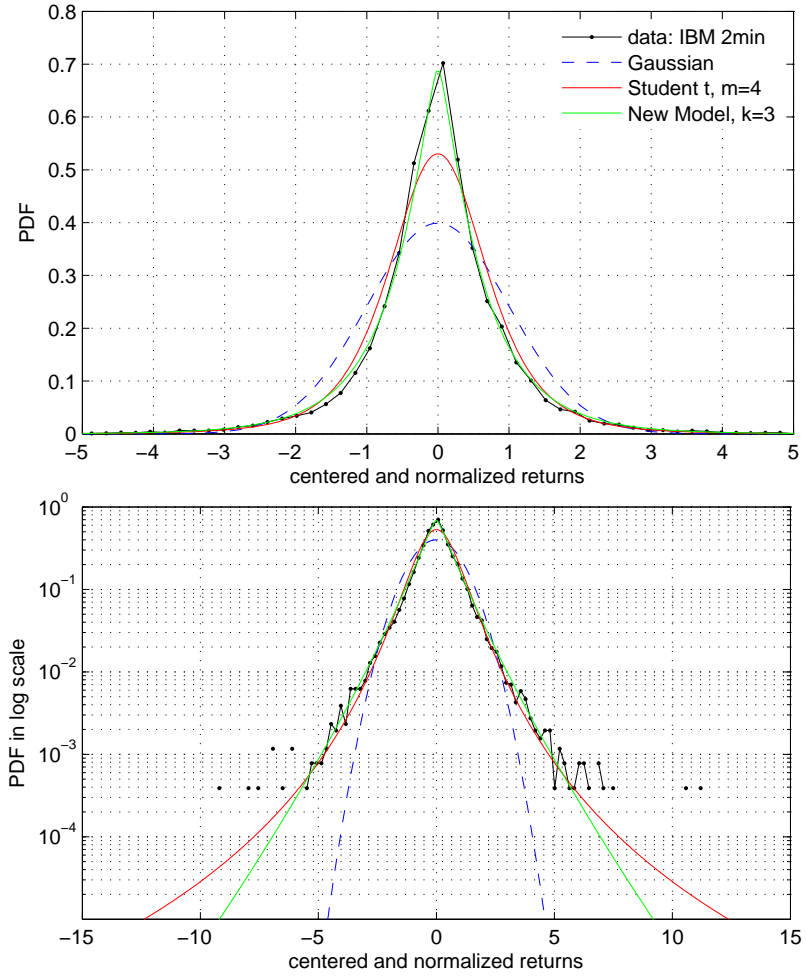


Figure 1.5: PDF of the centered and normalized IBM two-minute returns: the data set is TAQ from Jun 20th 2007 to Sep 20th 2007.

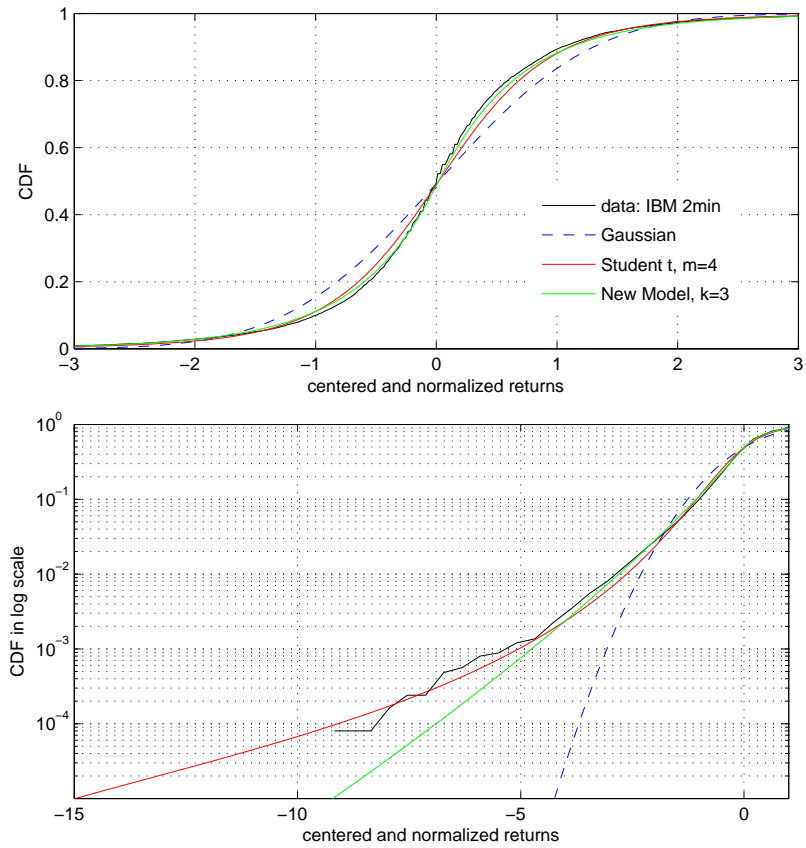


Figure 1.6: CDF of the centered and normalized IBM two-minute returns: the data set is TAQ from Jun 20th 2007 to Sep 20th 2007.

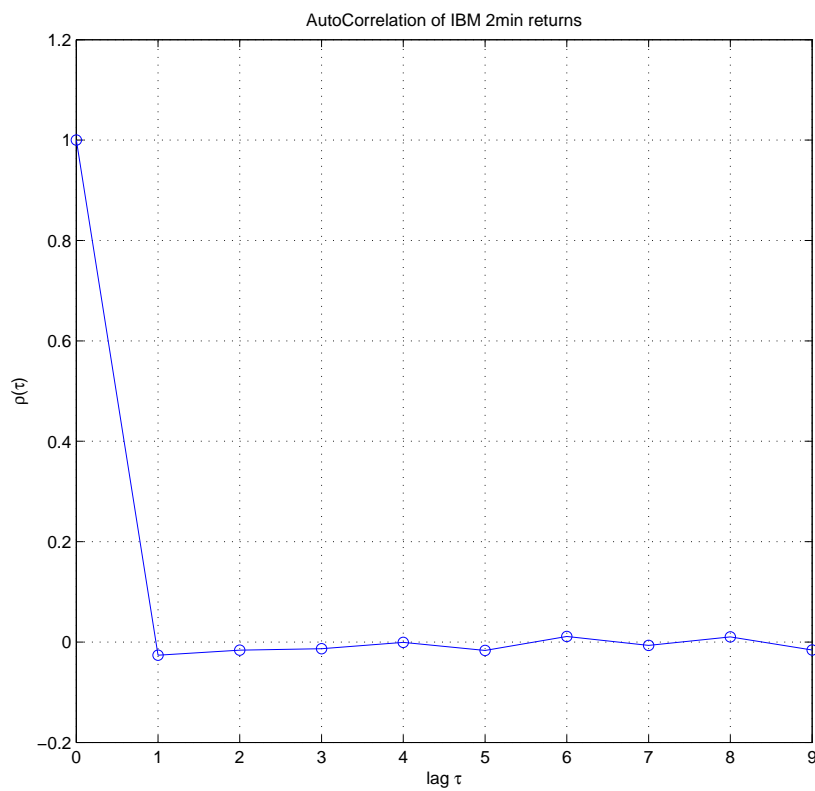


Figure 1.7: Autocorrelation of IBM two-minute returns: the data set is TAQ from Jun 20th 2007 to Sep 20th 2007.

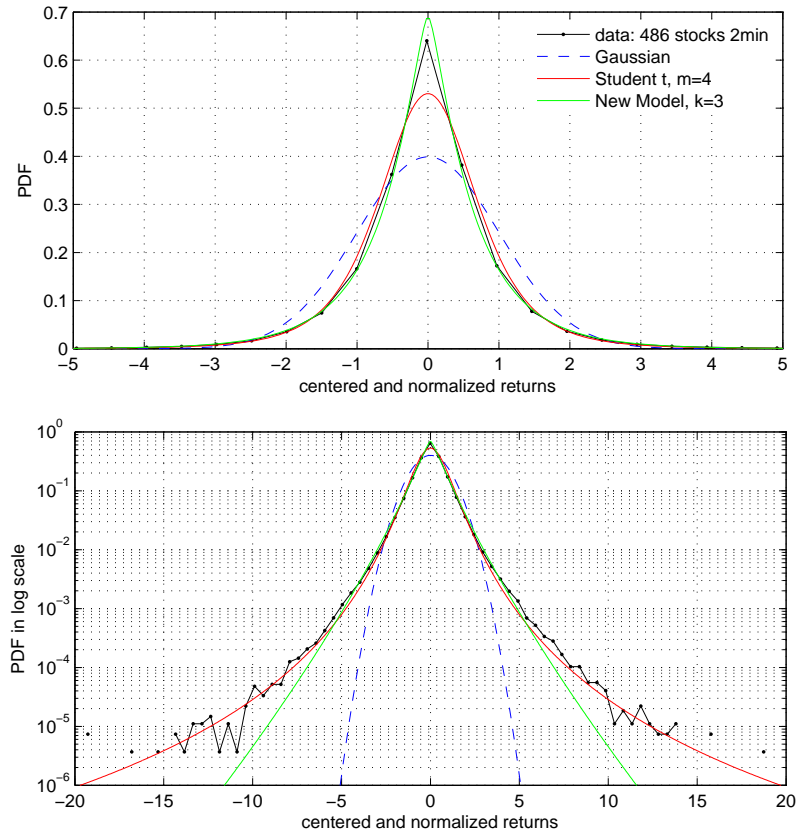


Figure 1.8: PDF of the centered and normalized two-minute returns for 486 stocks in *S&P500*: the data set is TAQ from Jun 20th 2007 to Sep 20th 2007.

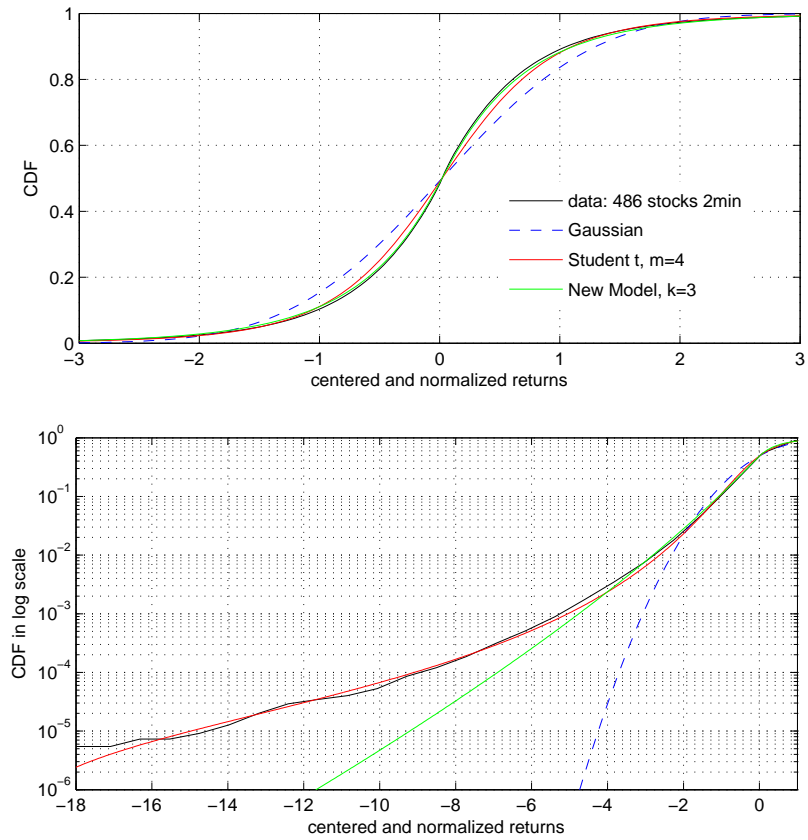


Figure 1.9: CDF of the centered and normalized two-minute returns for 486 stocks in *S&P500*: the data set is TAQ from Jun 20th 2007 to Sep 20th 2007.

### 1.2.2 Empirical distribution of multiple assets

As we explained in Section 1.1.3, the natural extension of the new model to the joint distribution of multi-asset returns has the feature that the probability density can blow up at the origin. The joint density of all assets having very small returns maybe much higher than it could be with a density bounded at the origin. Empirical data for short time returns of US equities are consistent with densities that blow up at the origin and are inconsistent with densities bounded at the origin. This is a significant form of dependence between returns on distinct assets that is not captured by return covariances. Indeed, even uncorrelated assets can and do have joint PDF's that blow up at the origin. If uncorrelated assets were independent (or if the principal components of correlated assets were independent) then the joint density would be bounded at the origin if the single asset marginals are bounded.

In order to show the unboundedness of the joint distribution of multi-asset returns, we design a test according to Theorem 1. Suppose we have  $p$  stocks, we denote the  $L_2$  norm of the centered and normalized return vector  $R$  to be  $z$ ,  $z = |R|$ ,  $R = [R_1, R_2, \dots, R_p]$ . We can find the empirical value of the following two quantities

$$y = \log(P(z < \epsilon)), \quad x = \log(\epsilon).$$

for small positive  $\epsilon$ . Then, we use the standard linear regression to find the coefficients  $a$  and  $b$  for

$$y = ax + b.$$

For example, we take two-minute returns of 5 stocks, IBM, DIS, XOM, MRK, KO. The two-minute returns are computed from the mid prices of TAQ data from 06/20/2007 to 09/20/2007 (64 business days and 195 data points per day) as are in Section 1.2.1. The number of observation is  $n = 195 \times 64 = 12480$  for every stock. We find the linear fitting coefficient  $a = 2.78$  with the standard error 0.04. We also study daily returns of the same 5 stocks, IBM, DIS, XOM, MRK, KO. The daily returns are computed using the adjusted close prices in Yahoo! Finance from 01/02/1970 to 06/02/2011. The number of observation is  $n = 10453$  for every stock. We find the linear fitting coefficient  $a = 2.65$  with the standard error 0.02.

By Theorem 1, the coefficient  $a$  is equal to  $k$  for the new model if  $k < p$ . The coefficient  $a$  is equal to the number of assets  $p$  as long as the PDF is continuous and non-zero at the mean. The examples includes the Gaussian, the Student t distributions and the new model with  $k > p$ . Thus, empirical evidence seems to show an unbounded probability density.

We did some numerical experiments to test the accuracy of the estimate of  $a$  above. For each of the three models (Gaussian, Student t with  $m = 4$ , the new model with  $k = 2$  and  $k = 3$ ) we created artificial datasets with the same sample size  $n$  and estimated  $a$  as above. We did this 5000 times independently for each of the four models. The result from the two-minute return data, presented in Figure 1.10, show that it is very unlikely to estimate  $a$  as low as 2.78 when the true probability density is bounded at the origin. Similarly, the result from the daily return data, presented in Figure 1.11, show that it is very unlikely to estimate  $a$  as low as 2.65 when the true probability density is bounded at the origin. The linear fitting coefficient  $a$  from the return data falls into the regime that the new model with  $k = 2, 3$  can

produce while it is out of the 99.9% confidence interval of those independent estimations from Gaussian and Student t.

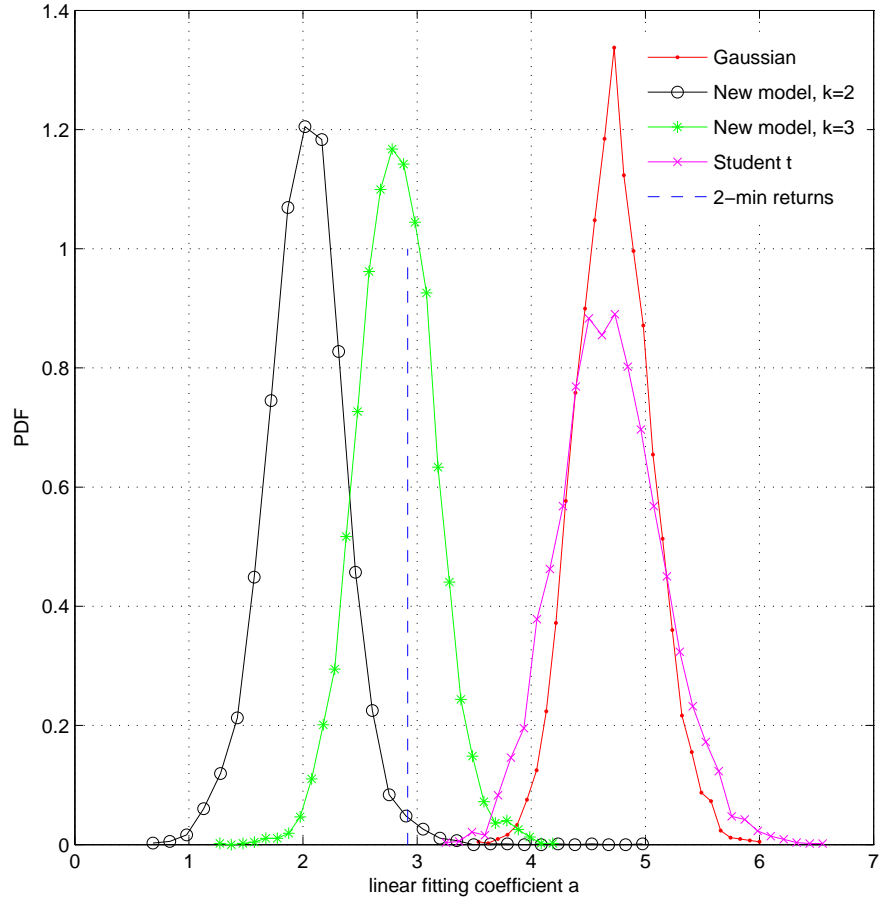


Figure 1.10: Compare the PDFs for the linear fitting coefficient  $a$  from the random samples to that from the two-minute returns of 5 stocks, IBM, DIS, XOM, MRK, KO. The number of observation is  $n = 195 \times 64 = 12480$ . The data set is TAQ from 06/20/2007 to 09/20/2007.



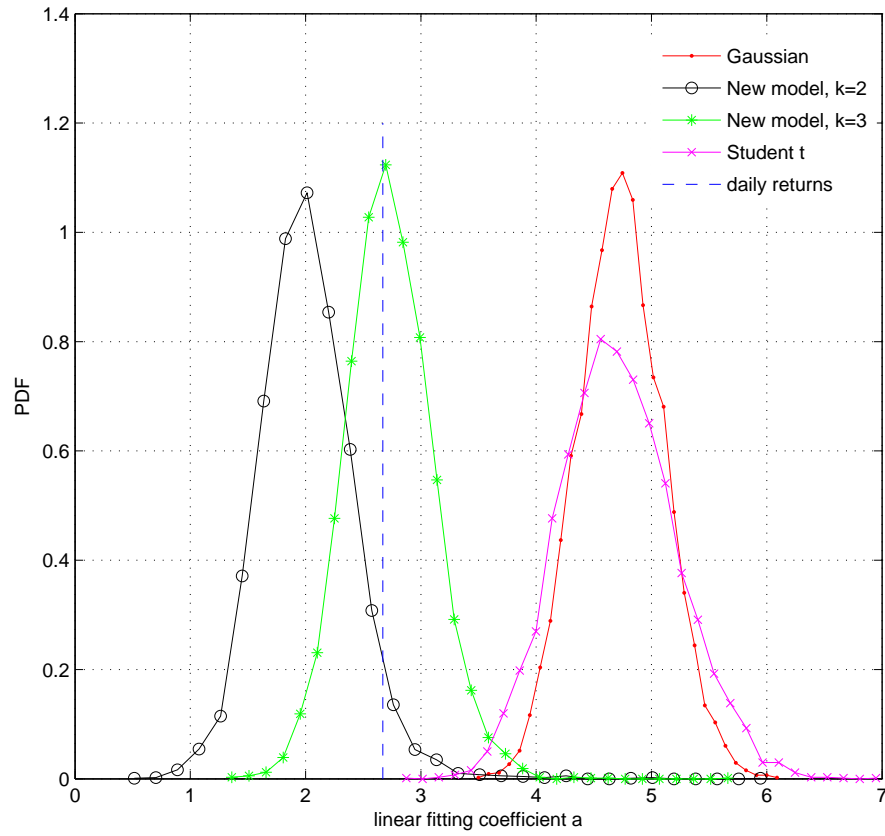


Figure 1.11: Compare the PDFs for the linear fitting coefficient  $a$  from the random samples to that from the daily returns of 5 stocks, IBM, DIS, XOM, MRK, KO. The number of observation is  $n = 10453$ . The data set is TAQ from 01/02/1970 to 06/02/2011.