Estimation and Calibration

7.1 Introduction

Virtually all models in finance have parameters that must be specified in order to completely describe them. Statistical estimation can be used for some or all of these parameters under two important conditions:

1. We believe that the parameter values do not change quickly in time.
2. We have access to sufficient (recent) historical data on the underlying process to provide reliable estimates of these parameter values.

Statistical estimation involves the distribution of the data and uses methods such as maximum likelihood estimation or least squares.

There are also some instances in which estimation of parameters is undesirable. For example we have seen that a risk-neutral distribution Q, used in the pricing of options, is not necessarily identical to the historical distribution P. Estimating parameter values for the distribution P based on historical data may have little or nothing to do with their values under Q. However, market information, for example the prices of options with a given underlying stock carry information about the expected value of certain functions of the stock under Q and the process of arranging parameter values so that the model prescribes options prices as close as possible to those observed is called *calibration* of the model.

Although there is a substantial philosophical difference between calibration and statistical estimation, there are many similarities in the methodology, since both often require obtaining the root of one or more functions or maximizing or minimizing a function and so these are the problems that we begin this section with.

There are many numerical problems that become much more difficult in the presence of noise, from finding the roots of a given function to solving a system of differential equations. Problems if this sort have given rise to substantial research in a variety of areas including stochastic approximation and Markov Chain Monte Carlo. In statistics, for example, we typically estimate parameters using least squares or maximum likelihood estimation. However, how do we carry out maximum likelihood estimation in cases where no simple form for the likelihood function is available? Suppose we are able to simulate data from a model with a given set of parameter values, say, but unable to express the likelihood function in a simple form amenable to numerical maximization.

For some concreteness, suppose we are interested in calibrating a model for the price of options on the SPX, the S&P500 index. The call option prices in Table 7.1 were observed on May 6, 2004 when the SPX was quoted at 1116.84. All options expired on September 18, corresponding to $T = .3699$ years. The option price is calculated as midpoint of the bid and ask, and K is the exercise price, $P_O(K)$, the corresponding option price.

K	950	1005	1050	1100	1125	1150	1175	1200	1225	1250
$P_O(K)$	173.1	124.8	88.5	53.7	39.3	27.3	17.85	10.95	6.2	3.25

Table 7.1. Price of SPX Call Options

Suppose we use the Black-Scholes formula to determine the implied volatility for these options. The implied volatility is the value of σ solving

$$BS(1116.84, K, r, 0.3699, \sigma) = P_O(K) \tag{7.1}$$

where r is the spot interest rate. If exact prices were known (rather than just the bid and ask prices), we could determine r from the put call parity relation since there are also put options listed with the same exercise price. However, here we used the rate $r = .01$ of a short-term treasury bill since it is difficult to find nearly simultaneous trades in put and call options with the same strike price.. The MATLAB function BLSIMPV(S_0,K,r,T,$P_O(K)$) returns the implied volatility i.e. provides the solution of (7.1). In Figure 7.1 we give what is often referred to as the volatility smile, although smirks and frowns are probably more common than smiles these days (surely a sign of the times). This is a graph of the exercise price of the option against the implied volatility of the option. The fact that this implied volatility is not constant is further evidence that the Black-Scholes model (at least with constant parameters) does not fit the risk-neutral distribution very well.

We saw in Chapter 3 that distributions such as the normal inverse Gaussian distribution can do a better job of fitting historical distributions of stocks and indices than does the normal and so we might hope that it can be applied successfully to the fit of a risk-neutral distribution. The main problem in applying distributions such as the NIG is that of calibrating the parameters of the model to market data. In models such as this, if option prices are obtained from a simulation, how can we hope to obtain parameter values for the model that are consistent with a set of market prices for the options? More specifically suppose we assume that the price of a stock at time T is given by

$$S_T = S_0 e^{rT} \exp(X)$$

where X has the NIG$(\alpha, \beta, \delta, \mu)$ distribution of Lemma 29. In order that the discounted future price forms a martingale, we require that $E(\exp(X)) = 1$ and

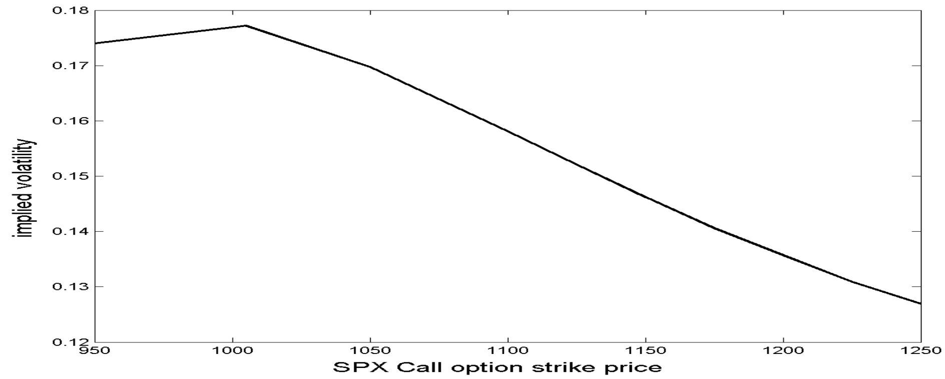

Figure 7.1: The volatility “smile” for SPX call options

from (3.23) with $s = 1$ this implies that

$$\mu = \delta\{\sqrt{\alpha^2 - (1+\beta)^2} - \sqrt{\alpha^2 - \beta^2}\} + \frac{1}{4}\ln\left(\frac{\alpha^2 - \beta^2}{\alpha^2 - (1+\beta)^2}\right)$$

and

$$|\beta + 1| < \alpha.$$

One of the four parameters, μ, is determined by the martingale condition so the other three can be used to fit to market option data if we wish. The price of a call option with exercise price K is now

$$E(S_0 \exp(X) - e^{-rT}K)^+$$

and this expectation is either a rather complicated numerical exercise or one that can be estimated by simulation. Whether or not we choose to price this option using simulation, the calibration problem reduces to selecting values of the parameters so that the theoretical option price agrees with the market price, i.e.

$$E(S_0 \exp(X) - e^{-rT}K)^+ = P_0(K). \tag{7.2}$$

If we have 10 options as in Table 7.1 and only three parameters to vary, there is no hope for exact equality in 7.2 (10 equations in 3 unknowns is not promising). We could either fix two of the parameters, e.g. α, β and select the third, δ to agree with the price of a specific option or we could calibrate all three parameters to three or more options with similar strike prices and minimize the sum of the squared differences between observed and theoretical option prices. We will return to this problem later, but for the present notice that in order to solve it, we need to be able to either find roots or minima of functions when evaluations of these functions are noisy and have measurement or simulation error superimposed. We begin a discussion of this problem in the next section.

7.2 Finding a Root

Let us begin with the simple problem of identifying the root of a non-decreasing function, i.e. solving for the value x_0 satisfying

$$f(x_0) = 0.$$

The standard algorithm for this is the Newton-Raphson algorithm which begins with an arbitrary value of x_1 and then iterates

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{7.3}$$

Suppose, however, that our observations on the function $f(x)$ are of the form

$$Y(x) = f(x) + \varepsilon_x$$

where the errors ε_x at the point x are small, with mean 0 and variance σ_x^2. Even if we know the gradient $D = f'(x_0)$ of the function at or near the root (whether we use the gradient at the root or at the value x_n approaching the root does not materially effect our argument), the attempt at a Newton-Raphson algorithm

$$x_{n+1} = x_n - \frac{Y_{x_n}}{D} = x_n - \frac{f(x_n)}{D} - \frac{\varepsilon_{x_n}}{D} \tag{7.4}$$

has a serious problem. If there is no error whatsoever in the evaluation of the function, the iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{D}$$

typically converges to the root x_0. However in (7.4), on each iteration we add to the "correct" Newton-Raphson iterant $x_n - \frac{f(x_n)}{D}$ an error term $\frac{\varepsilon_{x_n}}{D}$ which is often independent of the previous errors having variance $D^{-2}\sigma_{x_i}^2$.Therefore after a total of N iterations of the process, we have accumulated an error which has variance

$$D^{-2}\sum_{i=1}^{N}\sigma_{x_i}^2.$$

The fact that this error typically increases without bound means that continuing the iteration of (7.4) does not provide a consistent estimator of the root but one whose error eventually grows. Newton's method is inappropriate for any model in which errors in the evaluation of the function, when accumulated until convergence is declared, are comparable in size to the tolerance of the algorithm. In general as long as the standard deviation of the function evaluations σ_x is of the same order of magnitude as

$$\text{tolerance} \times D/\sqrt{N}$$

where N is the number of iterations before convergence was declared, then the accumulation of the error in (7.4) clouds any possible conclusion concerning the location of the root. Suppose for example you declare a tolerance of 10^{-4} and it takes about 100 iterations for your program to claim convergence. Assume for the sake of argument that $D = 0.01, N \simeq 100$. Then an error in the function evaluation of order $\sigma_x \simeq 10^{-5}$ will invalidate conclusions from the Newton-Raphson iteration.

There is a relatively simple fix to the usual Newton-Raphson method which at least corrects the lack of convergence, unfortunately at the price of greater insensitivity when we are further from the root. Suppose we multiply the constant D^{-1} by a sequence of constants a_n that depend on n so the iteration is of the form

$$x_{n+1} = x_n - a_n D^{-1} Y_{x_n}. \tag{7.5}$$

It is easy to see that as long as σ_x is bounded, for example and

$$\sum a_n^2 < \infty \tag{7.6}$$

then, under mild conditions on the function $f(x)$,then x_n converges. This is intuitively clear because the variance of the error beyond the N'th term is roughly $\sum_{n=N}^{\infty} a_n^2 \sigma_x^2$ and this approaches zero as $N \to \infty$. Furthermore, provided that

$$\sum a_n = \infty, \tag{7.7}$$

convergence is to a root of the function $f(x)$. This is the result of Robbins-Monro (1951). Having the correct or asymptotically correct gradient D in (7.5) really only matters asymptotically since it controls the rate of convergence once we are in a small neighbourhood around the root. What is more important and sometimes difficult to achieve is arranging that the sequence a_n is large when we are quite far from the root, but small (something like $1/n$) when we are very close.

The ideal sequence of constants a_n in (7.5) is approximately 1 (this is the unmodified Newton-Raphson) until we are in the vicinity of the root and then slowly decreases to $1/n$. Provided we had some easy device for determining whether we are near the root, the Robbins-Monro method gives an adequate solution to the problem of finding roots to functions when only one root exists. However, for functions with many roots, other methods are required.

While searching for a root of a function, we are inevitably required to estimate the slope of the function at least in the vicinity of the current iterant x_n, because without knowledge of the slope, we do not know in what direction or how far to travel. The independent values of Y_{x_n} in the Robbins-Monro algorithm (7.5) do not serve this purpose very well. We saw in Chapter 4 that when estimating a difference, common random numbers provide considerable variance reduction. If the function $f(x)$ is evaluated by simulation with a single uniform input U to the simulation, we can express f in the form

$$f(x) = \int_0^1 H(x, u) du$$

for some function H. Alternatively

$$f(x) = E\{\frac{1}{n}\sum_{i=1}^{n} H(x, U_i)\} \tag{7.8}$$

where U_i are the pseudo-random inputs to the i'th simulation, which again we assume to be uniform[0,1]. There are many ways of incorporating variance reduction in the Monte Carlo evaluation of (7.8). For a probability density function $g(z)$ other than the $U[0,1]$ probability density function, we may wish to re-express (7.8) using importance sampling as

$$f(x) = E\{\frac{1}{n}\sum_{i=1}^{n} \frac{H(x, Z_i)}{g(Z_i)}\} \tag{7.9}$$

where the inputs Z_i are now distributed according to the p.d.f. $g(z)$ on [0,1]. Ideally the importance density $g(z)$ is approximately proportional to the function $H(x_0, z)$ where x_0 is the root and so we may wish to adjust g as our iteration proceeds and we get closer to the root and the shape of the function $H(x, z)$ changes. Since the expectation is unknown we could attempt to find the roots of an approximation to this, namely,

$$\widehat{f_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{H(x, Z_i)}{g(Z_i)}. \tag{7.10}$$

Then $\widehat{f_n}(x) \to f(x)$ as $n \to \infty$. Since the error in this approximation approaches zero as $n \to \infty$, it seems reasonable to apply Newton's method to $\widehat{f_n}(x)$,

$$\begin{aligned} x_{n+1} &= x_n - \frac{\widehat{f}_n(x_n)}{\widehat{f}'_n(x_n)} \\ &= x_n - \frac{\sum_{i=1}^{n} w_i H(x_n, Z_i)}{\sum_{i=1}^{n} w_i \frac{\partial}{\partial x} H(x_n, Z_i)} \end{aligned} \tag{7.11}$$

with weights

$$w_i = \frac{1}{g(Z_i)}.$$

Notice that even if the function $H(x, Z_i)$ is only known up to a multiplicative constant, since this constant disappears from the expression (7.11), we can still carry out the iteration. Similarly, the weights w_i need only be known up to a multiplicative constant. There are two primary advantages to this over the simplest version of the Robbins-Monro. The gradient has been estimated using common random numbers Z_i and we can therefore expect it to be a better estimator. The ability to incorporate an importance sampling density $g(z)$ also provides some potential improvement in efficiency. There is an obvious disadvantage to any method such as this or the Newton-Raphson which require the derivatives of the function $\frac{\partial}{\partial x}H(x_n, Z_i)$. If the function $f(x)$ is expressed as an integral against some density other than the uniform, for example if

$$f(x) = \int_{-\infty}^{\infty} H(x,w)h(w)dw$$

for density $h(w)$ then (7.11) requires a minimal change to

$$\begin{aligned} x_{n+1} &= x_n - \frac{\sum_{i=1}^{n} w_i H(x_n, Z_i)}{\sum_{i=1}^{n} w_i \frac{\partial}{\partial x} H(x_n, Z_i)} \text{ where} \\ w_i &= \frac{h(Z_i)}{g(Z_i)}. \end{aligned}$$

Let us now return to the calibration of the Normal Inverse Gaussian model to the option prices in Table 7.1. We will attempt a calibration of all the parameters in Chapter 8 but for the moment, for simplicity we will reuse the parameter values we estimated in Chapter 1 for the S7P500 index, $\alpha = 95.23$, $\beta = -4.72$. Since μ is determined by the martingale condition this leaves only one parameter δ, the analogue of variance, to calibrate to the option price. We wish to solve for δ, with $S_0 = 1116.84$, $r = 0.01, T = .3699$, $\alpha = 95.23$, $\beta = -4.72$ and μ determined by the martingale condition

$$\int_{-\infty}^{\infty} (S_0 \exp(x) - e^{-rT}K)^{+} nig(x; \alpha, \beta, \delta, \mu)dx = P_0(K)$$

where nig is the normal inverse Gaussian probability density function.

Since we are seeking the root over the value of δ, the derivative of the density $nig(x; \alpha, \beta, \delta, \mu)$ is required for the gradient and this is cumbersome (again involving Bessel functions) so we replaced it by fitting a cubic spline to the observations in a neighbourhood around δ and then using the derivative of this cubic spline. In Figure 7.2 we plot the values of $\sqrt{\delta}$ against the corresponding strike price K. We use the square root of δ since this is the closest analogue of the standard deviation of returns. The figure is similar to the volatility smile using the Black Scholes model in Figure 7.1 and, at least in this case, shows no more tendency for constancy of parameters than does the Normal model. If the NIG model if preferable for valuing options on the S&P500 index, it doesn't show in the calibration of the parameter δ to observed option prices.

Similar methods for finding roots can be used when the function f is a function of d arguments, a function from $\mathcal{R}^d$ to $\mathcal{R}^d$ and so both x_n and Y_{x_n} are $d-$dimensional but be forewarned, the problem is considerably more difficult in d dimensions when d is large. The analogous procedure to the Robbins-Monro method is

$$x_{n+1} = x_n - a_n D_n^{-1} Y_{x_n}$$

where the $d \times d$ matrix D_n is an approximation to the gradient of the function f at the root x_0. Once again, since we do not have precise measurements of

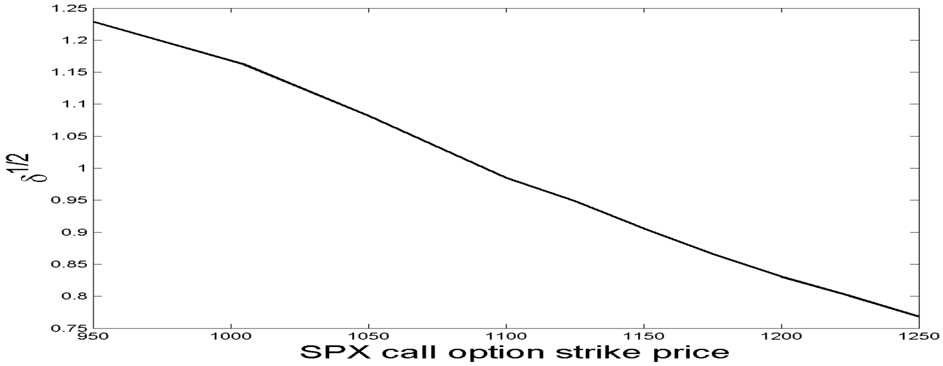

Figure 7.2: “Volatility” Smile for NIG fit to call option prices on the S&P500.

the function f, finding an approximation to the gradient D_n is non-trivial but even more important than in the one-dimensional case, since it determines the direction in which we search for the root. For noisy data, there are few simple procedures that give reliable estimates of the gradient. One possibility that seems to work satisfactorily is to fit a smooth function or spline to a set of values of (x, Y_x) which are in some small neighbourhood of the current iterant x_n and use the gradient of this function in place of D_n. This set of values (x, Y_x) can include previously sampled values $(x_j, Y_{x_j}), j < n$ and/or new values sampled expressly to estimate the gradient.

7.3 Maximization of Functions

We can use the results of the previous section to maximize a unimodal function $f(x)$ provided at on each iteration we are able to approximate the gradient of the function with a noisy observation

$$Y_x \simeq f'(x)$$

by simply using (7.5) to solve the equation $f'(x) = 0$.

For example suppose we can express the function $f(x)$ as

$$f(x) = \int_0^1 H(x,u)du$$

for some function H so that

$$f(x) = E\{\frac{1}{n}\sum_{i=1}^{n} H(x, U_i)\} \tag{7.12}$$

Again, the U_i are the pseudo-random inputs to the i'th simulation, which we can assume to be uniform[0,1]. For a probability density function $g(z)$ we can rewrite (7.12) using importance sampling as

$$f(x) = E\{\frac{1}{n}\sum_{i=1}^{n}\frac{H(x,Z_i)}{g(Z_i)}\} \tag{7.13}$$

where the inputs Z_i are distributed according to the p.d.f. $g(z)$ on [0,1]. We have many alternatives to finding its maximum if the function is smooth. For example we may use the Robbins-Monro method discussed in the previous section to find a root of $\widehat{f'}(x) = 0$ where

$$\widehat{f'}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{\frac{\partial}{\partial x}H(x,Z_i)}{g(Z_i)}.$$

Again there is a lack of consistency unless we permit the number of simulations n to grow to infinity, and our choice of ideal importance distribution will change as we iterate toward convergence. With this in mind, suppose we assume that the importance distribution comes from a one-parameter family $g(z;\theta)$ and we update the value of this parameter from time to time with information obtained about the function $H(x,z)$. In other words, assume that θ_n is a function of $(Z_1,...,Z_n)$ and $Z_{n+1}|\theta_n$ is drawn from the probability density function

$$g(z;\theta_n).$$

Then the sequence of approximations

$$\widehat{f_n}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{H(x,Z_i)}{g(Z_i;\theta_{i-1})}$$

all have expected value $f(x)$ and under mild conditions on the variance of the terms $var(\frac{H(x,Z_i)}{g(Z_i;\theta_{i-1})}|Z_1,...,Z_{i-1})$, then $\widehat{f_n}(x) \to f(x)$ as $n \to \infty$. Since the error in this approximation approaches zero as $n \to \infty$, it seems reasonable to apply Newton's method to such a system:

$$x_{n+1} = x_n - \frac{\widehat{f'_n}(x_n)}{\widehat{f''_n}(x_n)} \tag{7.14}$$

$$= x_n - \frac{\sum_{i=1}^{n} w_i \frac{\partial}{\partial x}H(x_n,Z_i)}{\sum_{i=1}^{n} w_i \frac{\partial^2}{\partial x^2}H(x_n,Z_i)} \tag{7.15}$$

with weights

$$w_i = \frac{1}{g(Z_i;\theta_{i-1})}.$$

Notice that even if the function $H(x,Z_i)$ is only known up to a multiplicative constant, since this constant disappears from the expression (7.15), we can still carry out the iteration. Similarly, the weights w_i need only be known up to a multiplicative constant.

Evidently there are at least three alternatives for maximization in the presence of noise:

1. Update the sample Z_i on each iteration and use a Robbins-Monro procedure, iterating
$$x_{n+1} = x_n - a_n \frac{\partial}{\partial x} H(x_n, Z_n)$$
2. Fix the number of simulations n and the importance distribution and maximize the approximation as in (Geyer, 1995)
$$x_n = \arg\max\{\frac{1}{n}\sum_{i=1}^{n} \frac{H(x, Z_i)}{g(Z_i)}\}$$
3. Update the importance distribution in some fashion and iterate (7.15).

The last two methods both require an importance distribution. The success of the third alternative really depends largely on how well the importance distribution $g(Z_i)$ adapts to the shape of the function $H(x, Z_i)$ for x near the maximizing value.

7.3.1 Example: estimation with missing data.

This example is a special case of a maximization problem involving "Missing Data" or latent variables. Although this methodology arose largely in biostatistics where various types of censorship of data are common, it has application as well in many areas including Finance. For example, periods such as week-ends holidays or evenings or periods between trades for thinly traded stocks or bonds could be considered "missing data" in the sense that if trades had occurred in these periods, it would often simply the analysis.

For a simple example suppose we wish to the daily data for a given stock index such as the Toronto Stock Exchange 300 index, TSE300, but on a given day, July 1 for example, the exchange is closed. Should we simply assume that this day does not exist on the market and use the return over a two day period around this time as if it was the return over a single day? Since other indices such as the S&P500 are available on this day, it is preferable to analyze the data as a bivariate or multivariate series and essentially impute the "missing" values for the TSE300 using its relationship to a correlated index. The data is in Table 7.2.

Date (2003)	June 30	July 1	July 2	July 3	July 4	July 7	July 8
TSE300 Close	6983.1	*	6990.3	6999.8	7001.9	*	7089.6
S&P500 Close	974.5	982.3	993.8	985.7	*	1004.42	1007.84

Table 7.2 Closing values for the S&P500 and TSE300, July 2003.
"*"= missing value

We will give a more complete discussion of dealing with missing data later when we introduce the EM algorithm but for the present we give a brief discussion of how "imputed values", (a statisticians euphemism for educated guesses) can help maximize the likelihood in problems such as this one. So for the moment, suppose that X is the "complete" data set (in the above example this is the data with the missing values "*" filled in) and Z be the data we actually observe. Then since the probability density for X factors into the marginal density for Z and the conditional density for X given Z,

$$f_x(x) = f_z(z)f_{x|z}(x|z)$$

we have upon taking logarithms and then the conditional expectation given $Z = z$ the relation

$$\ln(f_z(z)) = E[\ln(f_x(X))|Z = z] - \int \ln(f(x|z))f(x|z)dx \qquad (7.16)$$

where the second term is the entropy in the conditional distribution $f(x|z)$. Now suppose that we were somehow able to randomly generate the missing observations from their correct distribution, perhaps many times, while leaving the observed data alone. Averaging the log likelihood $\ln(f_x(X))$ for the completed data sets while holding the observed data Z constant at z is equivalent to approximating $E[\ln(f_x(X))|Z = z]$.

The maximum likelihood estimator of the parameters based on the incomplete data Z is obtained by maximizing the incomplete data likelihood $\ln(f_z(z))$ over the value(s) of the parameters, and a simple attempt at this would be to maximize the first term on the right side of (7.16) instead. Of course there is no guarantee that this is equivalent to maximizing both terms, and we return to this question later, but for the moment suppose that our objective is maximization of

$$E[\ln(f_x(X))|Z = z]$$

over the parameter values. In the above example, we can assume that the daily returns for the two indices are correlated normally distributed random variables so that in order to model these indices we need the 2 mean values and the 3 parameters in the covariance matrix for a bivariate normal distribution. Our strategy will be to "fill in" or impute the missing values, evaluate the likelihood $\ln(f_x(X))$ as if these imputed observations were real data, and regard this as a "noisy" evaluation of the function $E[\ln(f_x(X))|Z = z]$ which we would like to maximize.

In order to put this data on the more familiar ground of the bivariate normal distribution, we fit the log-normal distribution so that (starting with $t = 0$ on June 30, 2003), the close of the TSE at time t is given by

$$6983.1 \exp\{\sum_{i=1}^{t} Y_i\}$$

and of the S&P500,

$$974.5 \exp\{\sum_{i=1}^{t} X_i\}$$

where the daily returns (X_i, Y_i) are correlated normal random variables with mean (μ_x, μ_y) and with covariance matrix

$$\Lambda = \begin{pmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}.$$

Then the conditional log-likelihood (given the values on June 30) for the data, assuming all of the returns are observed, is

$$l_C(\mu, \Lambda) = -\frac{n}{2}\ln(|\Lambda|) - \frac{1}{2}\Sigma_{i=1}^{n}\left(\begin{pmatrix} Y_i \\ X_i \end{pmatrix} - \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}\right)' \Lambda^{-1}\left(\begin{pmatrix} Y_i \\ X_i \end{pmatrix} - \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}\right). \tag{7.17}$$

We would like to evaluate this log-likelihood when the missing data is replaced by imputed data and then maximize over the parameters. To this end we need to know how to impute missing values, requiring their conditional distribution given what is observed. So let us return to a small portion of the data as shown in Table 7.3. Notice that we are essentially given the total return $S = 0.001$ for the TSE over the two-day period July 1&2 but not the individual returns Y_1, Y_2 which we need to impute.

Date (2003)	July 1	July 2	Total Return
TSE300 return (Y_i)	Y_1	Y_2	$S = 0.001$
S&P500 return (X_i)	0.008	0.0116	0.0196

TABLE 7.3 Returns for TSE300 and S&P500

In order to carry out the imputation for Y_1, Y_2, recall that bivariate normal data is often expressed in terms of a linear regression relationship describing the conditional distribution of Y given X of the form

$$Y_i = \mu_y + \beta_{y|x}(X_i - \mu_x) + \varepsilon_i$$

where the error term ε_i is a Normal$(0, (1-\rho^2)\sigma_y^2)$ random variable independent of X_i and the regression parameter $\beta_{y|x}$ is related to the covariance matrix by

$$\beta_{y|x} = \frac{cov(Y, X)}{var(X)} = \rho\frac{\sigma_y}{\sigma_x}.$$

It follows from Problem 1 at the end of the Chapter that given the values X_1, X_2 and the sum $S = Y_1 + Y_2$, the distribution of Y_1 is normal

$$N(\frac{S}{2} + \frac{\beta_{y|x}}{2}(X_1 - X_2), \frac{1}{2}(1-\rho^2)\sigma_y^2). \tag{7.18}$$

If we are given tentative values of the parameters $\beta_{y|x}, \rho, \sigma_y^2$, this distribution can be used to generate an imputed value of Y_1 with $Y_2 = S - Y_1$. Of course a similar strategy allows us to impute the missing values of X_i, X_{i+1} when Y_i, Y_{i+1} and $X_i + X_{i+1}$ are observed. In this case the value of X_i should be generated from a

$$N(\frac{X_i + X_{i+1}}{2} + \frac{\beta_{x|y}}{2}(Y_i - Y_{i+1}), \frac{1}{2}(1-\rho^2)\sigma_x^2) \tag{7.19}$$

distribution where

$$\beta_{x|y} = \rho\frac{\sigma_x}{\sigma_y}.$$

We are now ready to return to the maximization of the log-likelihood function

$$E[l_C(\mu,\Lambda)|\text{observed data}]$$

where $l_C(\mu,\Lambda)$ is defined at (7.17). The maximum likelihood estimator of μ, Λ from complete data is

$$\widehat{\Lambda} = \frac{1}{n}\Sigma_{i=1}^n \left(\left(\begin{array}{c} Y_i \\ X_i \end{array}\right) - \left(\begin{array}{c} \widehat{\mu_y} \\ \widehat{\mu_x} \end{array}\right)\right)\left(\left(\begin{array}{c} Y_i \\ X_i \end{array}\right) - \left(\begin{array}{c} \widehat{\mu_y} \\ \widehat{\mu_x} \end{array}\right)\right)' \tag{7.20}$$

$$\left(\begin{array}{c} \widehat{\mu_y} \\ \widehat{\mu_x} \end{array}\right) = \frac{1}{n}\Sigma_{i=1}^n \left(\begin{array}{c} Y_i \\ X_i \end{array}\right). \tag{7.21}$$

The imputed values can be treated as actual observations for the missing data provided that we impute them many times and average the results. In other words, the following algorithm can be used to estimate the parameters. We begin with some arbitrary scheme for imputing the missing values. For example we could replace the individual returns Y_1, Y_2 in Table 7.3 by their average $S/2 = 0.0005$. Then iterate the following

1. Treating the imputed values as if they were observed, use the complete data (observed and imputed) to obtain estimators $\widehat{\Lambda}$ and $\widehat{\mu_y}, \widehat{\mu_x}$.

2. Use the estimated parameters to estimate $\beta_{y|x}$ and $\beta_{x|y}$ and to again impute the missing X_i using (7.19) and the missing Y_i using the distribution (7.18).

3. Again use the complete data (observed and imputed) to obtain estimators $\widehat{\Lambda}$ and $\widehat{\mu_y}, \widehat{\mu_x}$. Use as our current estimate of these parameters the **average** of the values of $\widehat{\Lambda}$ and $\widehat{\mu_y}, \widehat{\mu_x}$ obtained on each iteration of steps 1 or 3.

4. Repeat steps 3 and 4 until the averages used as parameter estimators of μ and Λ appear to converge.

We ran through steps three and four above a total of $100,000$ times (in about 70 seconds cpu time on a laptop) and obtained estimates of the mean and covariance matrix by averaging these 100000 estimates, resulting in

$$\widehat{\Lambda}_{av} = \begin{pmatrix} 1.034 & -0.115 \\ -0.115 & 5.329 \end{pmatrix} \times 10^{-5} \tag{7.22}$$

$$\begin{pmatrix} \widehat{\mu_y} \\ \widehat{\mu_x} \end{pmatrix} = \begin{pmatrix} 0.0025 \\ 0.0056 \end{pmatrix}. \tag{7.23}$$

These can be compared with a naive estimator obtained by imputing the missing returns as the mean return over a two-day period, based on the returns in Table 7.4 in which the imputed returns are in bold letters:

Date (2003)	July 1	July 2	July 3	July 4	July 7	July 8
TSE300 (Expected) Return	**0.0005**	**0.0005**	0.0014	0.003	**0.0062**	**0.0062**
S&P500 (Expected) Return	0.008	0.0116	-0.0082	**0.0094**	**0.0094**	0.0034

Table 7.4: Returns and imputed returns for TSE300 and S&P500

These naive estimates are the mean and covariance matrix of the data in Table 7.4,

$$\widehat{\Lambda} = \begin{pmatrix} 1.5 & 1.8 \\ 1.8 & 7.6 \end{pmatrix} \times 10^{-5},$$

$$\begin{pmatrix} \widehat{\mu_y} \\ \widehat{\mu_x} \end{pmatrix} = \begin{pmatrix} 0.0025 \\ 0.0056 \end{pmatrix}.$$

Notice that the two estimates of the mean are identical. This is not accidental, because essentially this is a numerical or Monte Carlo version of the EM algorithm discussed in the next section, and at least for estimating the mean, replacing each unobserved by its conditional expectation given the observations results in the correct maximum likelihood estimator. The covariance estimates, on the other hand, are quite different. Often, a covariance matrix obtained by naïve imputation, replacing observations with an average as we did with the data in Table 7.4, results in underestimating variances because the variation of to the missing observation is understated by an average. Similarly estimators of covariance or correlation may be biased for the same reason. The estimator (7.22), though it has random noise in it, does not suffer from bias to the same extent, because it has does a better job of replicating the characteristics of the random observations.

7.4 Maximum Likelihood Estimation

7.4.1 Missing and Censored Data

It is common in finance and generally in science for data to be biased by underreporting, and various forms of censorship. Patients who react badly to a treatment may not return for a follow-up appointment and the data that would have been obtained on this follow-up appointment is "missing" or unavailable, and the event that such data is unobserved may depend on the value that it would have taken. Similarly, data on the time to failure of companies can be "censored" or only partially observed if there is a merger or take-over during the study, since, in this case, we know only that the failure time of the original firm exceeds some value. Other forms of data-modification which introduce bias are also common. There is reporting bias when mutual funds that have had a poor quarter or year are more likely to give prominence to their 3 or 5-year returns in advertising or neglect to reply to a question on performance over a particular period. Managers, stocks, and companies with particularly poor returns disappear from data-bases, exchanges, or companies for obvious reasons. We have already seen some techniques for adjusting for the latter "survivorship bias" but in general, successful treatment of "missing" or "censored" data requires some insight into the mechanism by which the data goes missing or is censored.

In order to give a simple example of survivorship or reporting bias, suppose that the losses $x_i, i = 1, ..., n$ on portfolios managed by n managers over a period of time are assumed normally distributed and for simplicity we assume that all have the same mean μ and standard deviation σ. For various reasons we expect that the probability that a manager is no longer with us at the end of this period of time, (and so the loss for the whole period is "missing") is a function of the loss that would have occurred had they persisted so that the probability that a given data value x_i is observed is $\pi(x_i)$, a non-increasing function of the value x_i. The probability that an observation x_i is missing is therefore $1 - \pi(x_i)$. Then the likelihood for this data is

$$L(\mu, \sigma) = \prod_{i=1}^{n-1} \{\pi(x_i) n(x_i; \mu, \sigma^2)\}^{\Delta_i} \{\int (1 - \pi(z)) n(z; \mu, \sigma^2) dz\}^{1-\Delta_i} \qquad (7.24)$$

where $\Delta_i = 1$ if the i'th value is observed and otherwise $\Delta_i = 0$ if it is missing, and $n(x_i; \mu, \sigma^2)$ denotes the normal(μ, σ^2) probability density function at the point x_i. The first term

$$\pi(x_i) n(x_i; \mu, \sigma^2)$$

is the likelihood corresponding to an event "the $i'th$ loss is x_i and it is observed" and the last term

$$\int (1 - \pi(z)) n(z; \mu, \sigma^2) dz = E[1 - \pi(X_i)]$$

is the probability that an observation is missing. Suppose $n = 7$ and all of the $x_i, i = 1, ..., n - 1$ are observed but the last value x_7 is missing. In this case we know the values of x_i for all but the last observation, which could be replaced by a randomly generated observation if we knew the parameters μ, σ and the function $\pi(x)$. If we wish to maximize the likelihood (7.24) then one simple approach is to complete the data by generating a missing or latent observations $X_7 = Z$ with the normal(μ, σ^2) distribution and then accepting the observation (as missing) with probability $1 - \pi(Z)$. The likelihood is proportional to

$$E\{H(\mu, \sigma, Z)\}$$

where the expectation is over the distribution of Z,

$$H(\mu,\sigma,Z)=H(\mu,\sigma,Z,\theta)=\frac{1}{\sigma^7}\exp\{-\frac{\sum_{i=1}^6(x_i-\mu)^2+(Z-\mu)^2}{2\sigma^2}\}\frac{(1-\pi(Z))}{g(Z;\theta)}$$

and we assume Z has probability density function

$$g(z;\theta).$$

Consider maximizing this function over μ in the special case that $\sigma=1$ and $\pi(z)=e^{-z},z>0$ and $g(x;\theta)$ is an exponential probability density function with mean θ. We might fix the number of simulations m and the importance distribution and maximize the approximation to the integral:

$$\mu_n=\arg\max\{\frac{1}{m}\sum_{i=1}^m H(\mu,Z_i)\}$$

This method recommended by Geyer(1996) and has the advantage that it essentially estimates the function unbiasedly and then one can rely on standard maximization programs such as *fminsearch* in Matlab. It requires however that we fix the importance distribution $g(z,\theta)$ and the number of simulations. Alternatives are Robbins-Monro methods which have the ability to update the importance distribution parameter θ to reduce the variance of the estimator.

Various stochastic volatility models and "Regime switching" models can also be viewed as examples of latent variables or missing data problems. If the market has two volatility states, "high" and "low" these can not be directly observed but only inferred from the movement in the market over a period of time. In this case, given the state that the market is in, the parameters are often easily estimated, but this state is a latent variable that needs to be estimated or imputed from the observations.

7.4.2 Example: Mixture Models

One common problem which can also be viewed as a missing data problem is that of estimation of parameters in a mixture model. To take a simple example, consider a vector of observations of stock returns $(x_1,...,x_n)$ and assume that x_i has a probability density function that takes one of two possible forms $f_1(x|\theta_1)$ or $f_2(x|\theta_2)$ where the two parameters θ_1,θ_2 are not specified. We assume that a given observation comes from population distribution $f_1(x|\theta_1)$ with some probability p_1 and otherwise, with probability $p_2=1-p_1$ it comes from the population with density $f_2(x|\theta_2)$. This is an example of a mixture of two distributions: the overall population distribution is

$$f(x)=p_1f_1(x|\theta_1)+p_2f_2(x|\theta_2).$$

If each data point came with an identifier z_i taking possible values $1,2$ that specified the parent distribution, estimation of the parameters p_1,θ_1,θ_2 would be easy. We would estimate p_1 using the proportion of observations from population 1 and then use these to estimate the parameter θ_1, and similarly θ_2. However, this identifier is not usually available. For example it is a common model in finance to assume that there are two categories of observations, or states of the market, corresponding to high volatility (θ_2 large) and low volatility (θ_1 small) but of course we can only infer which state we are currently in by the magnitude of the last few changes in price. Alternatively we might assume that the vector of identifiers $z_1, z_2, ...$ mentioned above forms a simple $2-$state Markov chain, in which case the model for stock returns is commonly referred to as a *regime-switching* model.

Once again for the sake of a concrete problem, we use the returns from the S&P500 index between January 1, 2003 and May 11, 2004 to estimate a mixture model. In particular suppose that the density function or a return is assumed to be a mixture of normal distributions with mean μ and variances $\theta_i, i = 1, 2$, i.e.

$$\begin{aligned} & p_1 n(x;\mu,\theta_1) + (1-p_1)n(x;\mu,\theta_2) \\ &= p_1 f_1(x|\theta_1) + (1-p_1) f_2(x|\theta_2). \end{aligned}$$

For the sake of identifiability of the parameters, we assume $\theta_1 < \theta_2$. Consider first the simple case in which we assume that the identifiers z_i are independent and observed. Then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f_1(x_i|\theta_1)^{\Delta_i} f_2(x_i|\theta_2)^{1-\Delta_i}$$

where $\Delta_i = I(z_i = 1)$. Since Δ_i is not observed, the estimating function in the missing data case is obtained by conditioning the "complete data" estimating function on the observed information, i.e.

$$\sum_{i=1}^{n} E[\Delta_i|x_1...x_n] \frac{\partial}{\partial\theta_1} \ln f_1(x_i|\theta_1) = 0$$

$$\sum_{i=1}^{n} E[1-\Delta_i|x_1...x_n] \frac{\partial}{\partial\theta_2} \ln f_2(x_i|\theta_2) = 0.$$

In the case of independent z_i the conditional probability that $\Delta_i = 1$ is given by

$$w_i = \frac{p_1 f_1(x_i|\theta_1)}{p_1 f_1(x_i|\theta_1) + p_2 f_2(x_i|\theta_2)}$$

and this results in the estimating functions to be solved for θ,

$$\sum_{i=1}^{n} w_i \frac{\partial}{\partial\theta_1} \ln f_1(x_i|\theta_1) = 0 \tag{7.25}$$

$$\sum_{i=1}^{n} (1-w_i) \frac{\partial}{\partial\theta_2} \ln f_2(x_i|\theta_2) = 0. \tag{7.26}$$

Of course typically p_1, p_2 are not known either and they also need to be estimated from the data concurrently with our estimation of θ. The estimating function can be derived in exactly the same manner; if the Δ_i were observed, then we would estimate the parameters p_1, p_2 using

$$p_1 = \frac{1}{n}\sum_{i=1}^{n} \Delta_i$$

which after conditioning is of the form

$$\begin{aligned} p_1 &= \frac{1}{n}\sum_{i=1}^{n} E[\Delta_i|x_i] = \frac{1}{n}\sum_{i=1}^{n} w_i \text{ and,} \qquad (7.27) \\ p_2 &= 1 - p_1. \end{aligned}$$

Typically a simple routine which begins with initial value of the parameters and then iteratively steps towards a solution of (7.25) and 7.27) provides at least slow convergence to a solution. There is no absolute guarantee that the solution corresponds to a local maximum of the likelihood function, but for the S&P500 data mentioned the likelihood appears to increase monotonically. Convergence is to the model

$$0.811N(\mu, 0.0085^2) + 0.189N(\mu, 0.015^2),$$

which corresponds to a $\sqrt{252 \times .0085^2}$ =14% annual volatility on 81% of the days and 24% volatility the other 19% of the days. In general we need to be concerned about whether the initial values of the parameters affect the result but in this case varying the initial values of p_1, θ_1, θ_2 did not appear to have any effect (except in the time to convergence). This simple algorithm, when applied to mixture problems, can more generally be quite sensitive to the starting values.

When we estimate parameters as we did in the mixture problem above, whether we use the EM algorithm or one of its many relatives, missing components of the score function are replaced by their conditional expectation given the observations. For estimation of the mean in a normal family, we replace missing observations by their conditional expectation. On the other hand, these imputed values typically have less variability than the actual data we were unable to observe, and so if our interest is in the variance, the substitution for non-observed data by a conditional expectation tends to reduce the sample variance. If we have many parameters that we wish to estimate, we can essentially replace the missing data by repeated draws from the appropriate conditional distribution and use these repeatedly to estimate the parameter values and their variances. If the missing data is randomly generated, it should have the same characteristics as the data it is replacing, including the higher moments. The next method is a Bayesian implementation of this idea, one of few excursions into Bayesian methodology in this text.

The *Data Augmentation algorithm* (see Tanner and Wong, 1987) is used in Bayesian problems when there is latent or missing data as in the above example.

A Bayesian begins with a prior distribution $\pi(\theta)$ assigned to the parameters being estimated. This distribution reflects our prior knowledge of the parameter values and their variability and there are many possible choices for it, but ideally it should not have too much influence on the conclusions since we always wish the data to do more talking than the Bayesian's preconceived ideas of where the parameters lie. We explain the algorithm in the context of the example above so x is the vector of observations, and the missing data is denoted by Δ. The posterior joint distribution of the parameter θ and the missing data Δ is the conditional distribution of θ and Δ given the data x and can be written

$$p(\theta, \Delta|x) = p(\theta|x)p(\Delta|\theta, x).$$

The idea behind the data augmentation algorithm is to iteratively draw (θ, Δ) from approximations to the joint distribution of θ, Δ given the observed data. Typically $p(\theta|x)$ is difficult to compute but the conditional densities such as $p(\theta|x, \Delta)$ and $p(\theta|x, \Delta)$ are much easier to deal with.

The algorithm iterates these two steps:

1. Draw Δ^* from the density $p(\Delta|\theta^*, x)$.

2. Draw θ^* from the density $p(\theta|x, \Delta^*)$.

Iteration continues until the joint distribution of the pair (θ^*, Δ^*) appears to stabilize in which case they are regarded as a draw from the joint distribution $p(\theta, \Delta|x)$. The method is, in fact, a special case of the "Gibbs sampler" to be discussed later and its convergence to the correct conditional distribution derived from the balance equations we give there. At this point, to provide just a little credibility, we outline an intuitive argument for a Gibbs sampler in this simple case.

Consider the following problem. We wish to generate random variables (θ, Δ) having some joint distribution $p(\theta, \Delta|x)$. To save a little notation let us abbreviate $p(.|x)$ to $p_x(.)$. Unfortunately, the joint conditional distribution $p_x(\theta, \Delta)$is difficult to deal with, but conditional distributions such as

$$p_x(\theta|\Delta) = \frac{p_x(\theta, \Delta)}{p_x(\Delta)} \text{ and}$$
$$p_x(\Delta|\theta) = \frac{p_x(\theta, \Delta)}{p_x(\theta)}$$

are easy to sample from. We proceed as follows. Start with an arbitrary value for θ_1 and then generate Δ_1 from the density $p_x(\Delta|\theta_1)$. Now we generate θ_2 from the probability density $p_x(\theta|\Delta_1)$ and then Δ_2 from the density $p_x(\Delta|\theta_2)$ and so on. It is easy to see that the pairs of observations $(\theta_1, \Delta_1), (\theta_2, \Delta_2), \ldots$ constitute a two dimensional Markov Chain with transition kernel $K(\theta_{t+1}, \Delta_{t+1}|\theta_t, \Delta_t) = p_x(\theta_{t+1}|\Delta_t)p_x(\Delta_{t+1}|\theta_{t+1})$ and it is our hope that the joint density $p_x(\theta, \Delta)$ is an invariant distribution for this chain. To verify this note that

$$\begin{aligned}
&\int\int p_x(\theta_{t+1}|\Delta)p_x(\Delta_{t+1}|\theta_{t+1})p_x(\theta,\Delta)d\theta d\Delta \\
&= \int\int p_x(\theta_{t+1}|\Delta)p_x(\Delta_{t+1}|\theta_{t+1})p_x(\theta,\Delta)d\theta d\Delta \\
&= p_x(\Delta_{t+1}|\theta_{t+1})\int p_x(\theta_{t+1}|\Delta)p_x(\Delta)d\Delta \\
&= p_x(\Delta_{t+1}|\theta_{t+1})p_x(\theta_{t+1}) = p_x(\theta_{t+1},\Delta_{t+1}).
\end{aligned}$$

Since this is the invariant distribution, if the chain is ergodic, it will converge in distribution to $p_x(\theta,\Delta)$. There is no certainty in this or any other Markov chain that we have reached equilibrium after a finite number of draws, although there are many techniques for trying to monitor whether the distribution continues to change. There are other, more elaborate, methods for estimating the parameters in mixture models. See for example Robert (1996).

The algorithm is clearer if we apply it to the Normal mixture model for the S&P500 data discussed above. In this case we start with initial guesses Δ_i^* at the values of Δ_i. For example we could assign the smallest returns (in absolute value) to population one so $\Delta_i^* = 1$ and the rest to population 2, $\Delta_i^* = 0$ with about equal numbers in the two camps. Then θ^* is drawn with density $p_x(\theta|\theta,\Delta^*)$. This is easy to do in this example because the vector Δ^* assigns each observation to one of the two populations. For simplicity suppose that we have subtracted the mean from the returns so that we can assume $\mu = 0$ and let the parameters θ_1, θ_2 denote the *variances* of the two normal distributions.

For example suppose that the prior distribution of θ_1 is proportional to θ_1^{-1}

$$\pi(\theta_1) \propto \theta_1^{-1}, \theta_1 > 0.$$

There is no such *probability* density function, but there is a measure with this density and it is often convenient to permit such so-called *improper prior distributions.* The prior $\pi(\theta_1)$ corresponds to assuming an improper uniform prior distribution for $\log(\theta_1)$ because if U is uniform on some set then $\theta_1 = e^U$ will have density function

$$\propto |\frac{dU}{d\theta_1}| = \theta_1^{-1} \text{ on the corresponding set.}$$

Then the posterior density of θ_1 is

$$\begin{aligned}
p(\theta_1|x,\Delta^*) &\propto \prod_i [n(x_i;0,\theta_1)]^{\Delta_i^*}\theta_1^{-1} \\
&\propto \theta_1^{-\nu_1/2-1}\exp\{-\frac{\sum \Delta_i^* {x_i}^2}{2\theta_1}\}
\end{aligned}$$

where $\nu_1 = \sum_i \Delta_i^*$ is the degrees of freedom for this sample. The posterior density of the reciprocal of the variance $\psi = \theta_1^{-1}$ is

$$\begin{aligned} p(\psi|x, \Delta^*) &\propto \psi^{\nu_1/2+1} \exp\{-\frac{\psi \sum \Delta_i^* {x_i}^2}{2}\} |\frac{d\theta_1}{d\psi}| \\ &\propto \psi^{\nu_1/2-1} \exp\{-\frac{\psi \sum \Delta_i^* {x_i}^2}{2}\} \end{aligned}$$

We identify this posterior distribution of θ_1^{-1} as the Gamma($\nu_1/2$,$2/ss_1^2$) distribution with

$$ss_1^2 = \sum_{i=1}^n \Delta_i^* {x_i}^2$$

and this means that

$$ss_1^2 \theta_1^{-1}$$

has a chi-squared distribution with $\nu_1/2$ degrees if freedom.

Similarly with two parameters θ_1, θ_2, we may assume the prior density of $(\theta_1, \theta_2\ )$ is proportional to

$$\theta_1^{-2} \theta_2^{-2}$$

resulting in a independent Gamma($\nu_1/2$,$2/ss_1^2$), Gamma($\nu_2/2$,$2/ss_2^2$) posterior distributions for θ_1^{-1} and θ_2^{-1} where $\nu_2 = \sum_i (1 - \Delta_i^*)$ and

$$ss_2^2 = \sum_{i=1}^n (1 - \Delta_i^*) x_i^2.$$

The parameter p_1 could also have a prior distribution assigned and a posterior calculated in a similar fashion (e.g. assume a (non-informative) prior distribution for p_1 with density function proportional to $[p(1-p)]^{-1/2}$, so the posterior distribution given the value of Δ is that of a Beta $(\nu_1 + \frac{1}{2}, \nu_2 + \frac{1}{2})$ distribution (see Box and Tiao, 1973, Section 1.3)). However, for comparison with the mixture model above we chose p_1 equal to its estimated value there, 0.81.

The algorithm therefore is as follows, beginning with $k = 0$ and an initial guess at Δ^*,

1. Calculate
   $\nu_1 = \sum_i \Delta_i^*$, $\quad ss_1^2 = \sum_{i=1}^n \Delta_i^* x_i^2$
   $\nu_2 = \sum_i (1 - \Delta_i^*)$ $\quad ss_2^2 = \sum_{i=1}^n (1 - \Delta_i^*) x_i^2$

2. Generate $\theta_1^{-1} \sim$ Gamma($\nu_1/2$,$2/ss_1^2$) and $\theta_2^{-1} \sim$Gamma($\nu_2/2$,$2/ss_2^2$) (and $p_1 \sim$Beta $(\nu_1 + \frac{1}{2}, \nu_2 + \frac{1}{2})$) if we wish to allow p_1 to vary.

3. Generate independent

$$\Delta_i^* \sim \text{Bernoulli}(\frac{p_1 f(x_i|\theta_1)}{p_1 f(x_i|\theta_1) + p_2 f(x_i|\theta_2)})$$

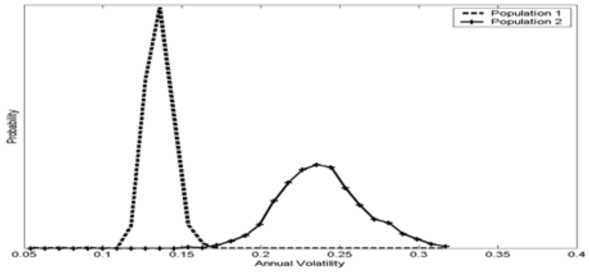

Figure 7.3: The posterior distribution of the annualized volatility parameters for the two components of a mixture model, fit to the S&P500 index, June-July 2003.

4. Return to step 1 until the distribution of $(\Delta_i^*, \theta_1^{-1}, \theta_1^{-1})$ has ceased to change. Then iterate steps 1-4 m times, considering these m subsequently generated values of $(\Delta_i^*, \theta_1^{-1}, \theta_1^{-1})$ as draws from the joint distribution of (Δ, θ) given $(x_1, ..., x_n)$.

The posterior distribution of the annualized volatilities determined by the data augmentation algorithm is give in Figure 7.3 and this gives a better idea of the error attached to the variances of the two normally distributed components for fixed p_1.

The next section provides a simple general technique for estimating parameters in the presence of missing or censored data.

7.4.3 The EM Algorithm

In principle maximum likelihood estimation is simple. We write down the likelihood or the probability of the observed values and then choose the parameter value which maximizes this expression. Difficulties occur primarily when the likelihood does not have a simple closed-form or easily evaluated expression. In the last section we saw an example in which we observed data Z but if complete data X were available (including "latent" or missing variables), then the complete-data likelihood takes a simple form as a function of the unknown parameter

$$L_C(\theta; x), \text{ } (C \text{ for "complete"}).$$

However each missing observation in a data set gives rise to a conditional expectation. If only a portion z of the complete data x is observed, then the likelihood function for the observed data, $L_O(\theta, z)$ say, ("O" is for "observed") is related to the complete data likelihood. As before, since

$$f_\theta(x) = f_\theta(z) f_\theta(x|z),$$

we have, for an arbitrary value of the parameter θ_0,

$$\begin{aligned}\ln(f_\theta(x)) &= \ln(f_\theta(z)) + \ln(f_\theta(x|z))\\ E_{\theta_0}[\ln(f_\theta(x))|z] &= E_{\theta_0}[\ln(f_\theta(z))|z] + E_{\theta_0}[\ln(f_\theta(x|z))|z]\\ &= \ln(f_\theta(z)) + \int \ln\{f_\theta(x|z)\} f_{\theta_0}(x|z)dx.\end{aligned}$$

Denote the term on the left side by

$$Q_{\theta_0}(\theta) = E_{\theta_0}[\ln(f_\theta(x))|z]. \tag{7.28}$$

Then the log-likelihood of the observed data $\ln\{L_O(\theta, z)\}$ is

$$\ln(f_\theta(z)) = Q_{\theta_0}(\theta) - \int \ln\{f_\theta(x|z)\} f_{\theta_0}(x|z)dx.$$

From this we obtain

$$\ln\{L_O(\theta, z)\} - \ln\{L_O(\theta_0, z)\} = Q_{\theta_0}(\theta) - Q_{\theta_0}(\theta_0) + H(\theta_0, \theta) \tag{7.30}$$

where

$$H(\theta_0, \theta) = E_{\theta_0}[\ln(\frac{f_{\theta_0}(x|z)}{f_\theta(x|z)}|z]$$

is the cross entropy between the conditional density of $x|z$ at the two values of the parameter.

The EM algorithm proceeds as follows: We begin with what we think is a good estimator of the parameter θ_0 and then maximize over the value of θ the function $Q_{\theta_0}(\theta)$. The maximizing value of θ, θ_1, say, is the next estimator. Replacing θ_0 by θ_1 we repeat, now maximizing $Q_{\theta_1}(\theta)$ and so on, obtaining a sequence of estimators $\theta_n, n = 1, 2, ...$ It is not hard to show that under fairly simple conditions, this sequence converges to the maximum likelihood estimator, i.e. the value $\widehat{\theta}$ satisfying $L_O(\widehat{\theta}; z) = \max\{L_O(\theta; z)\}$. In other words, the algorithm switches back and forth between the two steps starting with an initial guess at the parameter value θ_0 and $n = 0$ and stopping when the sequence θ_n appears to converge:

E Step Obtain the conditional expected value

$$Q_{\theta_n}(\theta) = E_{\theta_n}[\ln L_C(\theta; x)|z]$$

M Step Maximize this function $Q_{\theta_n}(\theta)$ over θ, letting

$$\theta_{n+1} = \arg\max Q_{\theta_n}(\theta)$$

be the next approximation to the parameter value. Increment n.

This algorithm works because of the identity (7.30). Notice that because of the M-step above, $Q_{\theta_n}(\theta_{n+1}) - Q_{\theta_n}(\theta_n) \geq 0$ and the term $H(\theta_{n+1}, \theta_n)$, because it is the cross-entropy between two distributions, is always non-negative. This shows that

$$\ln\{L_O(\theta_{n+1}, z)\} - \ln\{L_O(\theta_n, z)\} \geq 0$$

and that the observed data likelihood is a non-decreasing sequence. It must therefore converge.

The score function for the complete data

$$S_C(\theta; x) = \frac{\partial}{\partial\theta} \ln L_C(\theta; x)$$

is related to the observed data score $S_O(\theta; z) = \frac{\partial}{\partial\theta} \ln L_O(\theta; z)$ even more simply:

$$S_O(\theta; z) = E_\theta[S_C(\theta; x)|z]. \tag{7.31}$$

Therefore on the $E-$step of the EM algorithm, we may solve the equation for θ_{n+1} (checking as always that the solution corresponds to a maximum)

$$S_{obs}(\theta_{n+1}|\theta_n) = 0.$$

Thus, the EM algorithm can be more simply expressed as a simple resubstitution algorithm for solving the score equation (7.31) equals zero:

E Step For an initial guess at the parameter value θ_0, obtain the conditional expected value

$$S_O(\theta|\theta_n) = E_{\theta_n}[S_C(\theta; x)|z]. \tag{7.32}$$

Solve for θ_{n+1} the equation

$$S_O(\theta_{n+1}|\theta_n) = 0.$$

It is clear from this formulation that if the algorithm converges it must converge to a point θ satisfying

$$S_O(\theta|\theta) = 0$$

which, in view of (7.31) is the score equation for the observed data.

If there are m missing observations, notice that both (7.28) and (7.31) are $m-$fold integrals and may be difficult to compute. Nevertheless, (7.31) leads to an extremely useful observation that applies to missing or "censored" data, within the normal family of distributions and more generally, to exponential families. Equation (7.32) indicates that any term in the score function that is unobserved should be replaced by its conditional expectation given the observed data. If the score function contained terms like X^2 or $\sin(X)$ then the conditional expectation of such terms would need to be computed. This is, of course, NOT the same as replacing X by its conditional expectation and then using X^2 or $\sin(X)$. However, one of the marvelous features of the normal distributions is that, *for estimating the mean,* the score function is linear in X. For example, differentiating (7.17) with respect to μ_X, μ_Y we obtain the score function corresponding to estimation of μ for complete data.

$$\Sigma_{i=1}^{n}\Lambda^{-1}\left(\left(\begin{array}{c} Y_i \\ X_i \end{array}\right) - \left(\begin{array}{c} \mu_y \\ \mu_x \end{array}\right)\right)$$

Notice that taking the expected value given the observed data is equivalent to replacing each unknown component of X_i, Y_i by its conditional expectation given the observed. Moreover, within the multivariate normal family of distributions, taking conditional expectation is accomplished essentially by linear regression.

For a simple example of its application let us return to the problem with "missing" values in the S&P500 index and the TSE300 index discussed above. Suppose, for example we wish to fill in the value for the TSE300 on July 1 using conditional expectation. Which observations are directly relevant? Table 7.3 gives the daily returns for the two indices leaving out values that are independent of the unknowns Y_1, Y_2.

Recall from (7.18) that the conditional distribution of Y_i given X_i, X_{i+1} and $S = Y_i + Y_{i+1}$ is

$$N(\frac{S}{2} + \frac{\beta_{y|x}}{2}(X_i - X_{i+1}), \frac{1}{2}(1-\rho^2)\sigma_y^2). \tag{7.33}$$

and so the conditional expectation

$$E(Y_i|X_i, X_{i+1}, S] = \frac{S}{2} + \frac{\beta_{y|x}}{2}(X_i - X_{i+1})$$

with $\beta_{y|x} = \rho\sigma_y/\sigma_x$. Rather than naively replacing the unknown returns Y_i, Y_{i+1} by their average $S/2$ the term $\frac{1}{2}\beta_{y|x}(X_i - X_{i+1})$ provides an adjustment determined from the values of X. The regression coefficient $\beta_{y|x}$ can either be found by regressing Y on X preferably with data around the same period of time (since these coefficients tend to be somewhat time-dependent) or by using the current values of the maximum likelihood estimator of the covariance matrix Λ. Using the observed data for June and July 2003, we arrived at an estimated regression coefficient

$$\beta_{y|x} \simeq 0.432$$

and this allows to to fill in the missing values in the table with their conditional expected values:

Date (2003)	July 1	July 2
TSE300 Return	−0.0003	0.0013
S&P500 Return	0.008	0.0116

TABLE 7.4

Filling in the expected returns for all of the missing value for July 4 in the same way, but using the regression coefficient for X on Y of $\beta_{x|y} \simeq 1.045$, we observe the following for the unknowns in Table 7.5.

Date (2003)	July 1	July 2	July 3	July 4	July 7	July 8
TSE300 (Expected) Return	–0.0003	0.0013	0.0014	0.003	Y_7	Y_8
S&P500 (Expected) Return	0.008	0.0116	-0.0082	X_4	X_7	0.0034

TABLE 7.5

$$\begin{aligned}
X_4 &= \frac{1}{2}(0.0188 + 1.045(0.003 - Y_7)) \\
Y_7 &= \frac{1}{2}(0.0124 + 0.432(X_7 - 0.0034)) \\
X_4 + X_7 &= 0.0188 \\
Y_7 + Y_8 &= 0.0124
\end{aligned}$$

and these four equations in four unknown can be solved in the four unknowns giving

$$\begin{aligned}
Y_7 &= -0.0634, & Y_8 &= 0.0758, \\
X_4 &= 0.0441, & X_7 &= -0.0253.
\end{aligned}$$

Assuming that this pair of returns follows a correlated geometric Brownian motion then in order to estimate $cov(X,Y)$ in the presence of complete data we would solve the following equations:

$$\widehat{\mu_x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7.34}$$

$$\widehat{\mu_y} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7.35}$$

$$\widehat{cov(X,Y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu_x})(y_i - \widehat{\mu_y}). \tag{7.36}$$

For small sample size the estimator (7.36) can be adjusted for bias by replacing $\frac{1}{n}$ by $\frac{1}{n-1}$. The projection argument underlying the EM algorithm indicates that the maximum likelihood estimating function for the incomplete data is obtained by conditioning on the observed information. Projecting the three equations above on the observed data corresponds to doing the following: terms x_i, y_i or $(x_i - \widehat{\mu_x})(y_i - \widehat{\mu_y})$ in which one or both of (x_i, y_i) are unobserved,

1. If x_i is unobserved but $x_i + x_{i+1}, y_i, y_{i+1}$ are all observed, is observed, replace x_i by

   $$\widehat{x}_i = \frac{x_i + x_{i+1}}{2} + \frac{\beta_{x|y}}{2}(y_i - y_{i+1})$$

   and

   $$\widehat{x}_{i+1} = (x_i + x_{i+1}) - \widehat{x}_i.$$

2. If y_i is unobserved but $y_i + y_{i+1}, x_i, x_{i+1}$ are all observed, replace y_i by

$$\widehat{y_i} = \frac{y_i + y_{i+1}}{2} + \frac{\beta_{y|x}}{2}(x_i - x_{i+1})$$

and

$$\widehat{y}_{i+1} = (y_i + y_{i+1}) - \widehat{y}_i.$$

3. If all of x_i, y_i x_{i+1}, y_{i+1} are unobserved but $x_i + x_{i+1}$ and $y_i + y_{i+1}$ are observed, replace the two terms $(x_i - \widehat{\mu_x})(y_i - \widehat{\mu_y})$ and $(x_{i+1} - \widehat{\mu_x})(y_{i+1} - \widehat{\mu_y})$ in (7.36) by the single term

$$\frac{1}{2}(x_i + x_{i+1} - 2\widehat{\mu_x})(y_i + y_{i+1} - 2\widehat{\mu_y}).$$

Since there is now one fewer term in (7.36), reduce the value of n appearing there by one.

Here the regression coefficients $\beta_{x|y} = cov(x,y)/var(y)$ and $\beta_{y|x} = cov(x,y)/var(x)$ can either be estimated in a similar fashion from the data at hand, or estimated from a larger data-set in which both x_i and y_i are observed provided we think they are reasonably stable over time.

There are, of course, patterns for the missing data other than those considered in 1-3 above. They are rare in the case of two-dimensional data but much more likely in the case of higher dimensional correlated observations. The adjustments to the maximum likelihood estimators required to accommodate such patterns become increasingly complex, virtually impossible to provide a simple formula. The imputation method discussed earlier in this chapter, in which the missing observations are replaced not by their conditional mean but by imputed values with the same characteristics as the original complete data is the most successful method for providing estimates in such circumstances since it has only two ingredients, the relatively simple formulae for the complete data maximum likelihood estimators, and the ability to impute or generate the missing data with the correct distribution.

7.4.4 Monte Carlo Maximum Likelihood Estimation

Consider to begin with independent observations $X_1, ..., X_n$ all from the same probability density function $f_\theta(x)$. The maximum likelihood estimator of the parameter is the value θ which maximizes the likelihood

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

The ideal Monte Carlo methodology for maximum likelihood estimation would not require specification of the family of densities f_θ at all, but only the ability to generate variables X_i corresponding to any value of θ. However a moment's reflection will convince us that in general this is a tall order. By repeated simulation of variables X_i under the parameter value θ and averaging, Monte Carlo methods permit us to estimate unbiasedly any expected value, such as $E_\theta[\psi(X)]$ for arbitrary function ψ. Unfortunately the probability density function at a particular point x_i cannot be written as such an expected value, except in the case of a discrete random variable X where, of course, $P_\theta(X = x_i) = E_\theta[I(X = x_i)]$. If we have a method for estimating the density (or probability function if X is discrete) from simulations $X_i(\theta)$ taken for each parameter value θ, we might use these simulations to estimate $f_\theta(x)$ say with $\widehat{f}_\theta(x)$ and then maximize the estimated likelihood

$$\widehat{L}(\theta) = \prod_{i=1}^{n} \widehat{f}_\theta(x_i). \tag{7.37}$$

This estimator

$$\arg\max \prod_{i=1}^{n} \widehat{f}_\theta(x_i)$$

avoids any explicit form for the likelihood function at the cost of dealing only with an approximation to the likelihood and typically will require a very large simulation. Since the optimality properties of the maximum likelihood estimator depend on the local first and second derivatives of $f_\theta(x_i)$ in a neighbourhood of the true value it is essential that such an approximation be very accurate in a neighbourhood of the estimator. One of the most common methods for approximating $f_\theta(x)$ is that of kernel density estimation which relies on a kernel function $K(x)$ satisfying

$$\int_{-\infty}^{\infty} K(x)dx = 1,$$

often a probability density function centered around 0. Let $X_j(\theta), j = 1, ..., m$ denote simulated data under the parameter value θ, if possible using common random numbers. When the distributions have easily inverted cumulative distribution functions F_θ, for example, we could define $X_i(\theta) = F_\theta^{-1}(U_i)$ where U_i are independent uniform[0,1]. We must choose a window width parameter h controlling the degree of smoothing, and then we estimate the density using

$$\widehat{f}_\theta(x_i) = \frac{1}{mh} \sum_{j=1}^{m} K(\frac{x_i - X_j(\theta)}{h})$$

and so (7.37) becomes

$$\widehat{L}(\theta) = (mh)^{-n} \prod_{i=1}^{n} \{\sum_{j=1}^{m} K(\frac{x_i - X_j(\theta)}{h})\}. \tag{7.38}$$

Instead of operating completely free of the likelihood function, if we have partial information on f_θ, Monte Carlo Maximum Likelihood can be made more efficient. For example it is reasonably common to know f_θ only up to the normalizing constant (which, in theory we could obtain by laborious integration) and wish to estimate θ by maximum likelihood without evaluating this scaling constant. To be more specific, suppose that

$$f_\theta(x) = \frac{h_\theta(x)}{c(\theta)} \tag{7.39}$$

where the function $h_\theta(x)$ is known and relatively easily evaluated but $c(\theta)$ is unknown. There are many examples where probability density functions take a form such as this, but perhaps the most common is in a problem involving conditioning. For example suppose that we know the joint probability density function of two random variables $f_\theta(x,y)$ and we wish to condition on the value of one, say Y. Then the conditional probability density function is

$$f_\theta(x|y) = \frac{f_\theta(x,y)}{\int f_\theta(z,y)dz}$$

which is of the form (7.39) with $c(\theta)$ replaced by $\int f_\theta(z,y)dz$.

There are several methods available for generating a random sample from the density f_θ without using the constant $c(\theta)$ including acceptance-rejection and the Metropolis-Hastings algorithm. Notice that the likelihood for a sample of size n will take the form

$$\ln L(\theta) = \sum_{i=1}^n \ln h_\theta(x_i) - nE_\theta[\ln h_\theta(X)] \tag{7.40}$$

and so the constant is used to simply center the the estimating function $\sum_{i=1}^n \frac{\partial}{\partial\theta} \ln h_\theta(x_i)$ at its expected value. If we replace f_θ by a density with respect to some other value of the parameter or some completely different density function $g(x)$, then the maximizer of (7.40) satisfies

$$0 = \sum_{i=1}^n \frac{\partial}{\partial\theta} \ln f_\theta(x_i) \tag{7.41}$$

$$= \sum_{i=1}^n \frac{\partial}{\partial\theta} \ln h_\theta(x_i) - nE_\theta[\frac{\partial}{\partial\theta} \ln h_\theta(X)] \text{ or} \tag{7.42}$$

$$\frac{1}{n}\sum_{i=1}^n \frac{\frac{\partial}{\partial\theta} h_\theta(x_i)}{h_\theta(x_i)} = \ln E[\frac{\frac{\partial}{\partial\theta} h_\theta(X)}{g(X)}] \tag{7.43}$$

where the expected value in the second term is under the density $g(x)$ for X. Geyer(1996) suggests using this as the likelihood equation but with the expected value replaced by an average over Monte-Carlo simulations. In other words if we use N simulations of X_j generated under the density $g(x)$, we would solve for θ the estimating equation

$$\frac{1}{n}\sum_{i=1}^n \frac{\frac{\partial}{\partial\theta} h_\theta(x_i)}{h_\theta(x_i)} = \ln[\frac{1}{N}\sum_{j=1}^N \frac{h_\theta(X_j)}{g(X_j)}]. \tag{7.44}$$

This gives rise to several additional questions. How is the asymptotic normal distribution of the maximum likelihood estimator effected by the introduction of the Monte Carlo estimator of

$$\ln E[\frac{\frac{\partial}{\partial\theta}h_\theta(X)}{g(X)}],$$

and what sort of importance distribution $g(X)$ leads to the most efficient estimator of θ? For the first question, Geyer (1995) provides an estimator of the additional variance or covariance (over and above the usual covariance of a maximum likelihood estimator) introduced by the Monte Carlo estimation, and some discussion on the choice of importance distribution. The considerations around the choice of $g(x)$ are much the same here as in any application of importance sampling; we must avoid values of $g(x)$ which are too small and lead to large tails in the distribution of $\frac{\frac{\partial}{\partial\theta}h_\theta(X)}{g(X)}$. Mixtures are suggested, for example

$$g(x) = \sum_{j=1}^{k} p_j \frac{h_{\theta_j}(x)}{c(\theta_j)}$$

for various choices of the parameter values $\theta_1, ...\theta_k$ and accompanying weights $p_1 + ... + p_k = 1$. Unfortunately this introduces the additional parameters $p_1, ..., p_k$ which also need estimation, so it seems best to be parsimonious in our choice of k.

The rather restrictive condition that

$$f_\theta(x) = \frac{h_\theta(x)}{c(\theta)}$$

for known function $h_\theta(x)$ can be relaxed considerably above. In fact its primary purpose is to lead to the estimating function

$$\psi(x_i, \theta) = \frac{\frac{\partial}{\partial\theta}h_\theta(x_i)}{h_\theta(x_i)}$$

which was centered at its expectation using Monte Carlo and then added over all i to provide the estimating function for θ. The only purpose of the assumption (7.39) was to motivate a specific choice of function ψ. If we know virtually nothing about $f_\theta(x)$, but have the ability to generate random variables $X_i(\theta)$ for any given parameter value, we might well begin with a number of candidate estimating functions like $\psi_1(x,\theta), ...\psi_k(x,\theta)$, center them with Monte Carlo as before, and then determine the weighted average which results in the minimum possible asymptotic variance. When we do this, of course, we are not using the exact likelihood function (presumably this is unavailable), but since we are implicitly estimating the score function with a function of the form

$$\sum_i \beta_i(\theta)[\psi_i(x,\theta) - \mu_i(\theta)],$$

we are essentially approximating the likelihood with an exponential family distribution

$$c(\theta)\exp\{\sum_i B_i(\theta)\Psi_i(x,\theta)\}$$

with the functions Ψ_i either given by ψ_i or by $\frac{\partial}{\partial\theta}\psi_i$. Suppose we approximate the expectation

$$\mu_i(\theta) = E_\theta[\psi_i(X(\theta),\theta)]$$

using m simulations with

$$\widehat{\mu}_i(\theta) = \frac{1}{m}\sum_{j=1}^m \psi_i(X_j(\theta),\theta).$$

These same simulations allow us to estimate the covariance matrix $C_{ij}(\theta) = cov_\theta(\psi_i(X,\theta),\psi_j(X,\theta))$ as well as a difference of the form where h is small.

$$\widehat{\delta}_i(\theta) = \frac{1}{2mh}\sum_{j=1}^m [\psi_i(X_j(\theta+h),\theta) - \psi_i(X_j(\theta-h),\theta)]$$

(here it *is* important that we use common random numbers). Then the optimal estimating function constructed as a linear combination of the functions ψ_i is

$$\sum_j \beta_j(\theta)\sum_i[\psi_j(x_i,\theta) - \widehat{\mu}_j]$$

where the vector of values of β_j is estimated by

$$\widehat{\beta}^T(\theta) = \widehat{\delta}^T(\theta)C^{-1}(\theta). \tag{7.45}$$

In order assess the feasibility of this method, we consider a simple example, the symmetric stable laws whose density functions are quite difficult to evaluate, but for which it is easy to generate random variables. Suppose $X_1,...,X_n$ are assumed independent identically distributed with symmetric stable distributions, parameters $\mu = \theta$, with parameters $c = 1$ and α assumed known. Consider $M-$estimating functions of the form

$$\psi_k(x,\theta) = \sum_i \phi_k(x_i - \theta), k = 1,2,...4$$

where the candidate functions ϕ_k are

$$\phi_k(x) = \max(\min(x,\frac{k}{2}),-\frac{k}{2}), k = 1,...4.$$

Although we know that these functions satisfy

$$E_\theta[\phi_k(x_i - \theta)] = 0$$

for members of the symmetric stable family and therefore the functions $\psi_k(x,\theta)$ are unbiased, we do not wish to use that knowledge so we will continue to use Monte Carlo to generate the estimators $\widehat{\mu}_k(\theta)$. What combination of these four estimating functions provides the best estimator of θ? If we fix the value $c = 1$, we have seen that we can generate $X(\theta)$ using

$$X(\theta) = \theta + \sin(\alpha U) \left[\frac{\cos(U(1-\alpha))}{E} \right]^{\frac{1}{\alpha}-1} (\cos U)^{-1/\alpha} \tag{7.46}$$

with U_1 uniform$[-\pi/2, \pi/2]$ and E standard exponential, independent of U. A single pair (U, E) permits generating $X(\theta)$ for any θ. In this case, estimation of

$$\delta_k(\theta) = \frac{\partial}{\partial\theta}\psi_k(x,\theta) = \sum_i \phi_k'(x_i - \theta)$$

is a little easier since

$$\sum_i \phi_k'(x_i - \theta) = \sum_i I(|x_i - \theta| < k).$$

We tested this routine on data the data below, with $n = 21$, generated from a symmetric stable distribution with parameters $(0, 1, 1.5)$.

-3.1890	-2.0491	-1.8185	-1.7309	-1.5403	-1.3090	-1.2752
-0.8266	-0.7154	-0.2706	-0.2646	-0.0478	0.2827	0.9552
1.0775	1.1009	1.6693	1.6752	2.4415	2.4703	5.5746

Table 7.5: Data from the symmetric stable$(0, 1, 1.5)$ distribution

The mean of these values is 0.105 and the median is -0.2646. By running 1000 simulations with θ updated 100 times using a Robbins-Monro procedure and beginning with θ estimated by its median, we obtained an estimator of the parameter

$$\widehat{\theta} = -0.0954$$

with coefficients on the four estimating functions given by

$$\beta = (0.3561 \; 0.6554 \;\; 0.8705 \;\;\; 1.0096).$$

Evidently the last two functions $\psi_3(x,\theta), \psi_4(x,\theta)$ receive more weight than the first two. The resulting estimator is, of course, not really the maximum likelihood estimator, but it approximates the same to the extent that the functions ϕ_k and $\int \phi_k$ can be used to approximate the density function. Convergence of a Robbins-Monro algorithm here, as is frequently the case, is quite slow. To some extent this is a necessary payment made for our assumed ignorance about the form of the distribution and the expected value of the estimating functions. We are repeatedly approximating the latter using Monte Carlo and the noise so introduced results in the slow convergence. Here, as elsewhere, there is a price to pay for ignorance.

There are certainly more efficient estimators of the median of a symmetric stable distribution, but what is notable about this method is that it provides a "black-box" estimation of parameters requiring no knowledge of the distribution beyond the ability to simulate values with a particular parameter value and a set of vaguely reasonable estimating functions. Arguably, a fairly broad set of candidate estimating functions together with massive computer power may ultimately generate Monte Carlo estimators that rival maximum-likelihood estimation in efficiency without any knowledge of the structure of the distribution beyond the ability to generate samples from it.

7.5 Using Historical Data to estimate the parameters in Diffusion Models.

7.5.1 General Ito Processes

Typically a diffusion model for financial data includes parameters with unknown values which require estimation. We might, for example, wish to fit a diffusion model like

$$dr(t) = \mu(t, r(t))dt + \sigma(t, r(t))dW_t$$

to historical data on spot interest rates. Here as elsewhere we assume that the drift and diffusion coefficients are such that a solution exists. There are many choices of the drift and diffusion coefficients $\mu(t, r(t)), \sigma(t, r(t))$ leading to common models such as the Vasicek (1977), the CIR (Cox, Ingersoll, Ross, 1985), the Ho-Lee (Ho and Lee, 1986) model etc. but they all have unspecified parameters θ that must be either estimated or fit to price data before they can be used. Suppose for the present our intended use is *not* the pricing of bonds or interest rate derivatives, but using the model to simulate interest rate scenarios for an insurance company. In this case it is the P-measure that is our interest and so historical values of $r(t)$ are directly relevant. Unknown parameters θ may lie in either the drift term $\mu(t, r(t))$ or in the diffusion $\sigma(t, r(t))$ so we will add the argument θ to either or both function as required. There is a great deal of literature dealing with estimation of parameters in a diffusion model. We will largely follow McLeish and Kolkiewicz (1997).

According to the simplest discrete time approximation to the process, the *Euler Scheme*, the increments in the process over small intervals of time are approximately conditionally independent and normally distributed. Suppose we have already simulated the spot interest rate for integer multiples of Δt up to time $t = j\Delta t$. Provided the time increment Δt is small, if we denote $\Delta r = r(t + \Delta t) - r(t)$, we can generate the interest rate at time $t + \Delta t$ using the approximation:

$$\Delta r - \mu(t, r(t), \theta)\Delta t \sim N(0, \sigma^2(t, r(t), \theta))\Delta t). \tag{7.47}$$

Thus, if θ is a parameter in the drift term only, it can be estimated using weighted least squares; i.e. by minimizing the sum of the squared standardized normal variates.

$$min_\theta \sum_t w_t(\Delta r - \mu(t, r(t), \theta)\Delta t)^2$$

where the weights w_t are proportional to the *reciprocal of the variances* $w = 1/\sigma^2(t, r(t))$. The solution to this is obtained by setting the derivative with respect to θ equal to zero and solving for θ :

$$\sum_t w_t \frac{\partial\mu(t, r(t), \theta)}{\partial\theta}(\Delta r - \mu(t, r(t), \theta)\Delta t) = 0. \tag{7.48}$$

and it is not hard to see that this is the maximum likelihood estimator of θ assuming the normal approximation in (7.47) is exact. If the unknown parameter θ is in the diffusion term only, or in both drift and diffusion term, it is also easy to see that the log likelihood for the normal approximation is

$$-n\frac{1}{2}\ln(\sigma(t, r(t), \theta)) - \frac{1}{2}\sum_t \frac{(\Delta r - \mu(t, r(t), \theta)\Delta t)^2}{\sigma^2(t, r(t), \theta)}$$

with n the number of time points t at which we observed the spot interest rate. Maximizing this over the parameter θ results in the estimating function

$$\sum_t \frac{\partial\mu(t, r(t), \theta)}{\partial\theta}\frac{(\Delta r - \mu(t, r(t), \theta)\Delta t)}{\sigma^2(t, r(t), \theta)} + \frac{\partial\ln[\sigma^2(t, r(t), \theta)]}{\partial\theta}\{\frac{(\Delta r - \mu(t, r(t), \theta)\Delta t)^2}{2\sigma^2(t, r(t), \theta)} - 1\} = 0$$

which combines the estimating function for the drift (7.48) with that for the diffusion term, using weights inversely proportional to their variances.

Girsanov's Theorem allows us to construct maximum likelihood estimators also for the continuously observed processes analogous to some of the estimators above. For example, suppose the parameter θ resides in the drift term only, so the model is of the form

$$dX_t = \mu(t, X_t, \theta)dt + \sigma(t, X_t)dW_t, \quad X_o = x_0$$

Suppose P is the measure on $C[0, T]$ induced by X_t with $X_0 = 0$ and the measure P_0 on the same space is induced by a similar equation but with zero drift:

$$dX_t = \sigma(t, X_t)dW_t, \quad X_0 = x_0. \tag{7.49}$$

Then Girsanov's Theorem (see the appendix) asserts that with

$$M_t = \mathcal{E}(\int_0^t \frac{\mu(s, X_s, \theta)}{\sigma(s, X_s)}dW_s) = \exp\{\int_0^t \frac{\mu(s, X_s, \theta)}{\sigma^2(s, X_s)}dX_s - \frac{1}{2}\int_0^t \frac{\mu^2(t, X_t, \theta)}{\sigma^2(t, X_t)}ds\}, \tag{7.50}$$

under some boundedness conditions, then the Radon Nikodym derivative of P with respect to P_0 is given by M_T,

$$\frac{dP}{dP_0} = M_T.$$

The exponential martingale

$$M_t = \mathcal{E}(\int_0^t \frac{\mu(s, X_s, \theta)}{\sigma(s, X_s)} dW_s)$$

describes the Radon-Nikodym derivative for the processes restricted to $[0, t]$ and the process X_s appearing in (7.50) is assumed generated from the relation (7.49). Thus the maximum likelihood estimate of θ is obtained by maximizing M_T. Setting the derivative of its logarithm equal to 0 results in the likelihood equation

$$\int \frac{\partial\mu(t, X_t, \theta)}{\partial\theta} \frac{1}{\sigma^2(t, X_t)} dX_t - \int \frac{\partial\mu(t, X_t, \theta)}{\partial\theta} \frac{1}{\sigma^2(t, X_t)} \mu(t, X_t, \theta) dt = 0$$

$$\text{or} \quad \int \frac{\partial\mu(t, X_t, \theta)}{\partial\theta} \frac{1}{\sigma^2(t, X_t)} (dX_t - \mu(t, X_t, \theta)dt) = 0$$

Of course if we only had available observations taken at discrete time points $t_1 < t_2 < \ldots$ rather than the continuously observed process, we would need to replace the integral by a sum resulting in the same estimating function as (7.48), namely

$$\sum_{t_i} \sigma^{-2}(t_i, X_{t_i}) \frac{\partial\mu(t_i, X_{t_i}, \theta)}{\partial\theta} (\Delta X_{t_i} - \mu(t_i, X_{t_i}, \theta)\Delta t_i) = 0. \qquad (7.51)$$

For a continuous time model of the form

$$dX_t = \mu(t, X_t, \theta)dt + \sigma(t, X_t, \theta)dW_t$$

in which unknown parameters θ are in both the drift and the diffusion part of the stochastic differential, estimation of the parameter θ is very different than in the discrete time case. Consider, for example, solving the estimating equation

$$\psi(\theta) = \sum_i \{(\Delta X_{t_i} - \mu(t_i, X_{t_i}, \theta)\Delta t_i)^2 - \sigma^2(t_i, X_{ti}, \theta)\Delta t_i\} = 0 \qquad (7.52)$$

to obtain an estimator of θ. Since for sufficiently small time increments Δt_i the Euler increments are approximately normal, this estimating function is asymptotically unbiased as $\max\{\Delta t_i\} \to 0$. Moreover, for any normal random variable Z with mean 0 $var(Z^2) = 2(var(Z))^2$ so the variance of the estimating function above is, in the limit,

$$\begin{aligned} var(\psi(\theta)) &\to 2\sum_i \sigma^4(t_i, X_{ti}, \theta)(\Delta t)^2 \\ &\leq \max_i\{\Delta t_i\} \sum_i \sigma^4(t_i, X_{ti}, \theta)(\Delta t_i) \\ &\to 0 \end{aligned}$$

the convergence to zero since $\sum_i \sigma^4(t_i, X_{ti}, \theta)(\Delta t_i)$ converges to the integral

$$\int_0^T \sigma^4(t, X_t, \theta)dt$$

which is finite, provided for example that the function $\sigma(t, x, \theta)$ is bounded. In general the asymptotic variance of the estimator obtained by solving an equation of the form (7.52) is, under fairly modest regularity assumptions on the estimating function ψ, asymptotic to

$$\frac{var(\psi(\theta))}{E^2[\frac{\partial}{\partial\theta}\psi(\theta)]}. \tag{7.53}$$

It is easy to see that $E[\frac{\partial}{\partial\theta}\psi(\theta)]$is asymptotic to

$$-2\int \frac{\partial\sigma(t, X_t, \theta)}{\partial\theta}\sigma(t, X_t, \theta)dt$$

which is typically non-zero if σ depends on θ. Then (7.53) approaches 0 since the denominator is asymptotically nonzero. We conclude that the estimator of θ determined by the estimating function (7.52) approaches perfection (i.e. its variance approaches zero) as $\Delta t \to 0$ and this it true regardless of how long or how short the time interval is over which the process has been observed. Restated, this says that for a continuously observed process, *a parameter in the diffusion term (including the volatility parameter) can be perfectly estimated from an arbitrarily short period of observation.* Now if you think this is too good to be true, you are right, at least in a practical sense. In practice, volatility is by no means known or estimated precisely, because the diffusion model does not fit on the minute time scale necessary for the above argument to go through. Far from having nearly perfect knowledge of the volatility parameters, volatility appears to change rapidly on any time scale for which the diffusion model is considered a good fit. In the next section we discuss various estimates of volatility obtained from a discretely observed time (geometric) Brownian motion.

7.6 Estimating Volatility

In the last section, we saw that those parameters in the diffusion coefficient of a general diffusion were all too easily estimated. In fact from an arbitrarily short segment of the path, we can, at least in theory, obtain an estimator of the volatility which is exact (is unbiased and has zero variance) because of the infinite variation of a diffusion process in these arbitrarily short periods of time. Information for estimating the diffusion coefficient obtains much more rapidly than for the drift, and in this respect the continuous time processes are quite different than their discrete analogues. Two diffusions processes with different diffusion coefficients are mutually singular. Intuitively this means that the sample paths generated by these two measure lie in two disjoint subsets of $C[0, T]$

and so that we can theoretically determine from a single sample path the exact diffusion term. Of course to determine which set a path lies in, we need to estimate or examine the quadratic variation of the process, something which requires a perfectly graphed sample path and an infinitely powerful microscope. The real world is considerably different than this idealized situation for several reasons. First, we never observe a process in continuous time but only at discrete time points that may be close together. Second, diffusion processes fit data on security prices, interest rates and exchange rates only when viewed over a longer time intervals than a minute, hour, or day. Short-term behaviour is very different; for example they usually evolve through a series of jumps corresponding to trades of varying magnitudes and frequencies. Third, information obtained from the derivatives market can be used to obtain *implied volatilities* but these are not the same as the$P$-measure or historical volatilities in a discrete-time model. In theory, they should agree in the continuous time Black-Scholes model, since the risk neutral measure has the same diffusion coefficient as does the P measure, but this is of limited use since volatilities can change rapidly.

The volatility parameter is the single most important parameter for pricing short-term options and arguably the most important parameter for virtually all financial time series. In Figure 7.4 we plot the price of a call option on a stock currently trading at \$10. The strike price of the stock is \$12 and the time to maturity is 1/2 year. Notice that the option price ranges from close to 0 when $\sigma = 0.1$ to around \$1.40 when $\sigma = 0.7$ indicating that different estimates of volatility will have substantial impact on the option price. In fact over a large part of the range of this graph the relationship between option price and volatility is nearly linear.

Complicating the estimation of σ is its time-varying behaviour, with periods of persistent high volatility followed by periods of relatively lower volatility. Failure to accommodate the behaviour of the volatility process may result in highly misleading models, substantial pricing errors, and the catastrophic consequences that can accompany them. We have already distinguished between *estimation of volatility* based on historical data and the *calibration* of the volatility parameter to option price data. The latter is required if we are pricing derivatives since it chooses a value of the volatility parameter that is most consistent with the observed prices of options and futures. The former is necessary if we are interested in volatility for the purpose prediction, risk management, or scenario generation. The implied volatility is obtained from the price of one or more heavily-traded or benchmark derivatives sold on the open market having the same price process S_t for the underlying asset. We determine the volatility parameter which produces the market price of a given option. For example suppose an option with strike price K, maturity T, and initial value of the stock S_0 is traded on the market at a price given by V. Then, because the Black-Scholes price is increasing in σ, we may solve the equation

$$BS(S_0, K, r, T, \sigma) = V \tag{7.54}$$

for the implied volatility parameter σ. Because $BS(S_0, K, r, T, \sigma)$ is monotone

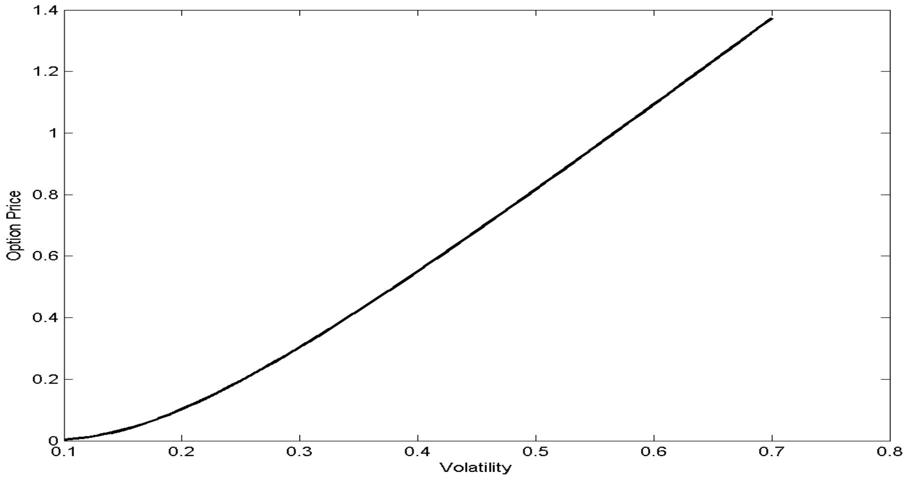

Figure 7.4: Price of Call option as a function of volatility, $S_0 = 10$, $K = 12, r = 0.05$, $T = 0.5$ years.

in σ a solution exists (as long is V is within a reasonable range) and is unique. This estimate of volatility differs from the historical volatility obtained by computing the sample variance of the returns $\log(S_{t+1}/S_t)$. Since it agrees with the market price of a specific option, it reflects more closely the risk-neutral distribution Q and is therefore used for pricing other derivatives. There are several disadvantages of this method of calibrating the parameter. The calibrated value of σ depends on the stock, the strike price of the option, as well as the current time t, the time to maturity $T - t$ and to some extent on other parameters that are harder to measure such as the liquidity of the option and the degree to which it is used as a benchmark. Nevertheless it is common to transform option prices V to volatility σ using (7.54) and then rather than deal with option prices, analyze these implied volatilities. This transformation may be justified if the resulting model is simpler or smoother as a function of σ. However there are many monotonic transformations of price that could do a better job of simplifying a non-Gaussian model.

In this section we will concentrate on the *estimation of volatility* based on historical data, as opposed to calibration. An estimator based on historical data approximates a parameter of the real world measure P whereas the calibrated parameter is a property of the risk-neutral measure Q. To explore the differences between implied and historical volatility, we downloaded two years of daily stock price data (Nov 1, 1998 to Nov 1, 2000) for Nortel (NT), listed on the New York Stock Exchange from the web-site http://finance.yahoo.com and on the basis of this, wish to estimate the volatility. We used the "adjusted

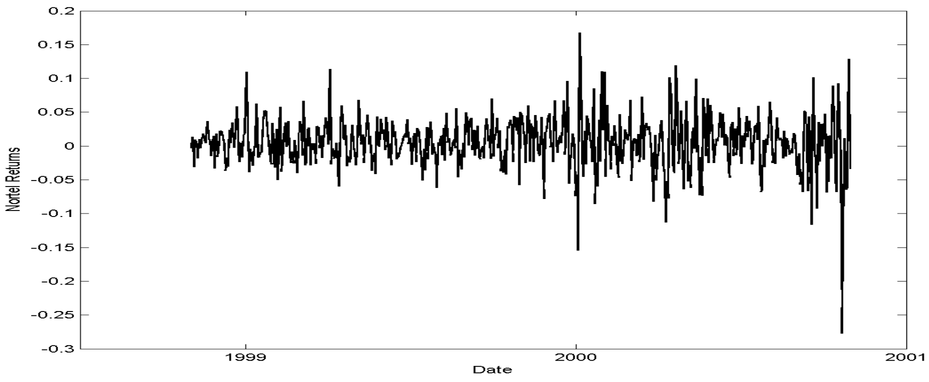

Figure 7.5: Nortel Returns, Nov 1, 1998-Nov 1, 2000.

close" since these have been adjusted for stock splits and dividends. The daily returns over this period are graphed in Figure 7.5 and the significant increase in volatility is evident towards the end of the sequence. Since the logarithm of daily stock prices is assumed to be a Brownian motion we may estimate the daily volatility using the sample variance of the first differences in these logarithms. To obtain the variance over a year, multiply by the number of trading days (around 253) in these years. Thus the annual volatility is estimated by sqrt(253*var(diff(log(ntprice)))) which gives a volatility of around 0.635.

How does this historical estimate of volatility compare with the volatility as determined by option prices?

Table 7.5 is a segment of a table of option prices obtained from the Chicago Board of Options Exchange and provides the current price (on November 2, 2000) of calls and puts on NT. There was a January $40 call selling for $\$8\frac{5}{8}$ and a $40 January put for $\$4\frac{1}{8}$. The price of the stock when these options prices were recorded was $44 (evidently the "good old days"!)

Calls	**Last Sale**	**Net**	**Bid**	**Ask**	**Volume**	**Open Interest**
01 Jan 40 (NT AH-E)	8 5/8	-7/8	8 3/8	8 7/8	79	10198
01 Jan 45 (NT-AI-E)	6 1/8	-1/4	5 7/8	6 1/4	460	4093
Puts	**Last Sale**	**Net**	**Bid**	**Ask**		
01 Jan 40 (NT MH-E)	4 1/8	+1/4	3 7/8	4 1/4		
01 Jan 45 (NT-MI-E)	6 3/8	+1/4	6 1/8	6 5/8		

Table 7.5: Option Prices for Nortel, Nov 2, 2000.

Suppose the current interest rate (in US dollars since these are US dollar prices) is 5.8%. This is roughly the interest rate on a short term risk free deposit like a treasury bill. Then the implied volatility is determined by finding the value of the parameter σ (it turns out to be around 0.79) so that the Black-Scholes formula gives exactly this value for an option price, i.e. finding a value of σ so that PUT=4.125 and CALL=8.625, for example:

[call,put]=blsprice(44,40,.058,55/253,.79,0) (returns call =8.6077, put =4.1066)

The implied volatility of a roughly at the money option is about $\sigma = 0.79$ and this roughly agrees with the price of the January 45 options

[call,put]=blsprice(44,45,.058,55/253,.79,0) (returns call =6.2453, put =6.6815)

Differences between the observed and theoretical option prices of the magnitude observed here are inevitable. Transaction cost and the fact that some options are more liquid than others can affect option prices. If we ignore this problem, we can find a pair of values (r, σ) by minimizing the sum of squared pricing error between the observed and Black-Scholes price of the put and call options, but because short-term option prices are quite insensitive to r, the estimates of r determined in this way can be unreliable.

We now have two estimates of volatility, historical volatility (0.635) and implied volatility (0.79). Which volatility is "correct"? The answer is both. The market conditions for this company changed enough over this period and subsequently to effect enormous changes in volatility and if we used a period of less than 2 years, our estimate of historical volatility would be larger (the one-year estimated volatility is 0.78, very close to the implied volatility). Moreover the implied volatility is a property of the Q measure, and this is determined not only by the recent history of the stock, but also investor risk preferences and fears. The Q measure is a distribution for stock prices assigned by *the market for options*. The P-measure was assigned by investor's past tastes for the stock, sometimes in a substantially different economic climate. It is often the case that "implied volatilities" are greater than the historical ones, although in theory for a Geometric Brownian motion process which allows frictionless continuous-time hedging, the risk-neutral Q and the actual distribution P should have the same volatilities.

7.6.1 Using Highs and Lows to estimate Volatility

There are many competing estimators of the volatility parameter of a geometric Brownian motion and to date we have used only one, the sample variance of stock returns. To set the stage for these estimators, suppose $S(t)$ is a geometric Brownian motion

$$d\ln(S(t)) = \mu dt + \sigma dW_t, S(0) = S_0$$

and we observe certain properties of non-overlapping segments of this process, for example days or weeks. In particular suppose $t_1 < t_2 < t_n..$ are the beginnings of certain observation periods of length Δ_i where $t_i + \Delta_i \leq t_{i+1}$ and so we can define, for this period, the open, the close, the high and the low respectively

$$\begin{aligned} O_i &= S(t_i), \quad C_i = S(t_i + \Delta_i) \\ H_i &= \max\{S(t); t_i \leq t \leq t_i + \Delta_i\}, \\ L_i &= \min\{S(t); t_i \leq t \leq t_i + \Delta_i\}. \end{aligned}$$

The observation periods may or may not be assumed equal in length (for example daily in which case Δ_i is around $\frac{1}{252}$ for all i). We will sometimes assume that the market does not move outside of these observation periods, in which case $O_i = C_{i-1}$ since in practice the open of many stocks at the beginning of the day is nearly identical to the close at the end of the previous day.

The simplest estimator of σ is that based on the variance of returns

$$var(R_i), \text{ where } R_i = \ln(C_i/O_i).$$

Since R_i is Normal($\mu\Delta_i$, $\sigma^2\Delta_i$), we can construct a regression problem towards estimating the parameters μ, σ

$$\frac{R_i}{\sqrt{\Delta_i}} = \mu\sqrt{\Delta_i} + \varepsilon_i, \text{ where } \varepsilon_i \text{ are i.i.d. } N(0, \sigma^2).$$

By regarding the left side as the response in an ordinary least squares regression, we obtain the standard estimators

$$\widehat{\mu} = \frac{\sum_{i=1}^n R_i}{\sum_{i=1}^n \Delta_i} \tag{7.55}$$

$$\widehat{\sigma}^2_{OC} = \frac{1}{n-1}\sum_{i=1}^n \frac{1}{\Delta_i}(R_i - \widehat{\mu}\Delta_i)^2, \tag{7.56}$$

both unbiased estimators of their respective parameters. Regression theory also provides the variance of these estimators:

$$var(\widehat{\mu}) = \frac{\sigma^2}{\sum_{i=1}^n \Delta_i},$$

$$var(\widehat{\sigma}^2_{OC}) = \frac{2\sigma^4}{n-1}.$$

There are more efficient estimators than $\widehat{\sigma}^2_{OC}$ if we include not only an open and a close but also the high and the low for those periods. In Chapter 5, we discovered that for a Geometric Brownian motion, the random variables $Z_{Hi} = \log(H_i/O_i)\log(H_i/C_i)$ and $Z_{Li} = \log(L_i/O_i)\log(L_i/C_i)$ are both exponentially distributed with expected value $\sigma^2\Delta_i/2$ and each of Z_{Li} and Z_{Hi} are independent of the open O_i and the close C_i. This auxiliary information can be used in addition to obtain an unbiased estimator of the parameter σ^2. For example consider the estimator

$$\widehat{\sigma}^2_{HH} = \frac{2}{n}\sum_{i=1}^n \frac{Z_{Hi}}{\Delta_i}$$

which, in view of the exponential distribution of Z_H is unbiased and has variance

$$var(\widehat{\sigma}^2_{HH}) = \frac{\sigma^4}{n},$$

about half of the variance based on returns. Of course we can build a similar estimator using the lows,

$$\widehat{\sigma}^2_{LL} = \frac{2}{n}\sum_{i=1}^{n}\frac{Z_{Li}}{\Delta_i}.$$

One might guess that an estimator which simply averages $\widehat{\sigma}^2_{HH}$ and $\widehat{\sigma}^2_{LL}$ would be four times as good as (7.56) but in fact because of the negative correlation between the two estimators being averaged it is better still! The average is the Rogers and Satchell (1991) estimator

$$\widehat{\sigma}^2_{RS} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_{Hi}+Z_{Li}}{\Delta_i} \tag{7.57}$$

which has variance depending on the correlation coefficient $\varsigma \simeq -0.338$ between Z_{Hi} and Z_{Li}

$$\begin{aligned} var(\widehat{\sigma}^2_{RS}) &= \frac{2}{n}var(\frac{Z_{Hi}}{\Delta_i})(1+\varsigma) \\ &\simeq \frac{\sigma^4}{2n}(1-0.338) \\ &\simeq 0.331\frac{\sigma^4}{n}. \end{aligned}$$

Evidently the Rogers and Satchell estimator is around 6 times as efficient (one sixth the variance) as the usual estimator or volatility (7.56). In fact it is independent of (7.56) as well and so we can combine the two with weights inversely proportional to the estimator variances to get a best linear unbiased estimator with weights rounded,

$$\widehat{\sigma}^2_{BLU} \simeq \frac{1}{7}\widehat{\sigma}^2_{OC} + \frac{6}{7}\widehat{\sigma}^2_{RS}$$

and this estimator has variance

$$var(\hat{\sigma}^2_{BLU}) \approx 0.284\frac{\sigma^4}{n}$$

around one seventh of (7.56). Using a single high, low, open, close for a seven day period to estimate volatility is roughly equivalent to using daily returns and the estimator (7.56). Related estimators have been suggested in the literature. See for example Parkinson(1980). Garman and Klass (1980) suggest the estimator

$$\hat{\sigma}^2_{GK} = \frac{1}{2}(\log(H_i/L_i))^2 - (2\ln(2)-1)(\log(C_i/O_i))^2 \tag{7.58}$$

which is similar to $\widehat{\sigma}^2_{BLU}$.

We could quibble over which of the above estimators to use, based on extraordinarily precise efficiency calculations for a geometric Brownian motion but much of this exercise would be wasted. Most indeed improve efficiency over (7.56) but not quite to the extent predicted by theory and provided that we use the extremes, the differences between those estimators which do use them are relatively small. To demonstrate this we compared them on the Dow Jones Industrial average for the period January 1999 to May 20, 2004. We computed all of the estimators for daily data on the open, close, high, and low and then plotted a 21-day moving average of these estimators against time. For example two of the estimators, $\widehat{\sigma}_{OC}$ and $\widehat{\sigma}_{HH}$ are given in Figure 7.8. The other estimators are not plotted since they closely resemble one of these two. The curve labelled "intraday volatility" measures the annualized volatility as determined by $\widehat{\sigma}_{HH}$ and that labelled "interday volatility", $\widehat{\sigma}_{OC}$. The estimators when evaluated on the whole data set provide values as follows:

$$\widehat{\sigma}_{OC} = 0.18 \quad \widehat{\sigma}_{HH} = 0.29 \quad \widehat{\sigma}_{LL} = 0.31 \quad \widehat{\sigma}_{RS} = 0.30 \quad \widehat{\sigma}_{BLU} = 0.29 \quad \widehat{\sigma}_{GK} = 0.31$$

indicating only minor differences between all of the estimators except $\widehat{\sigma}_{OC}$. Over this period January 1999 to May 20, 2004, the intraday volatility was consistently greater than the interday volatility, often by 50% or more, and there is a high correlation among the estimators that use the highs and lows as well as the open and close. This indicates a profound failure in the Geometric Brownian motion assumption for the Dow Jones Industrial index, and this discrepancy between two estimators of volatility is much greater than for any other index I have investigated. Since we have seen substantial differences in the variances of the estimators based on the geometric Brownian motion model, it is worth comparing the estimators for their empirical variance. However since the first, $\widehat{\sigma}_{OC}$ appears to have a different mean, we will compare them after normalizing by their mean: i.e. we compare values of

$$CV = \frac{sqrt(var(\widehat{\sigma}^2))}{E(\widehat{\sigma}^2))}$$

as estimated from the 21-day moving averages. These values are given below

$$CV_{OC} = 0.83 \quad CV_{HH} = 0.53 \quad CV_{LL} = 0.59 \quad CV_{RS} = 0.54 \quad CV_{BLU} = 0.55 \quad CV_{GK} = 0.55$$

and again, except for the first, they are virtually indistinguishable. What happened to the gains in efficiency we expected when we went from $\widehat{\sigma}_{HH}$ to $\widehat{\sigma}_{BLU}$? These gains are, in part at least, due to the negative correlation -0.338 between Z_{Hi} and Z_{Li} but for this data, they are positively correlated and in fact the 21-day moving averages of Z_{Hi} and Z_{Li} have correlation around 0.90. The only explanation that seems reasonable is that the DJA is quite far from a Geometric Brownian motion. If the volatility were stochastic or time-dependent, as most believe it is, then this can account for the apparently high correlation between Z_{Hi} and Z_{Li}. For example suppose that the volatility parameter σ_i was stochastic (but constant throughout the day) and possibly dependent on i.

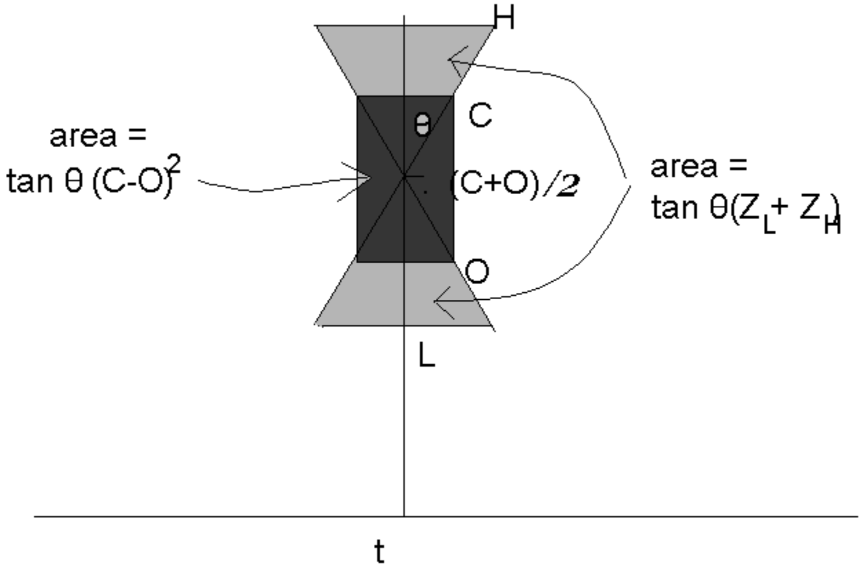

Figure 7.6: The construction of an Eggtimer plot

Then since

$$\begin{aligned} cov(Z_{Hi}, Z_{Li}) &= E[cov(Z_{Hi}, Z_{Li}|\sigma_i)] + cov(E[Z_{Li}|\sigma_i], E[Z_{Li}|\sigma_i]) \\ &= E[cov(Z_{Hi}, Z_{Li}|\sigma_i)] + var(\frac{\sigma_i^2 \Delta_i}{2}) \end{aligned}$$

it is possible for the first term to be negative (as predicted by a geometric Brownian motion model) and yet the second term is sufficiently large that the covariance itself is large and positive.

Thus we have three estimators of the volatility parameter,

$$\{\frac{\ln(C/O)\}^2}{\Delta t}, \frac{2\ln(L/O)\ln(L/C)}{\Delta t}, \frac{2\ln(L/O)\ln(L/C)}{\Delta t}\}.$$

While the first is independent of the other two given O, unfortunately the second and third are themselves not uncorrelated. In order to weight them optimally we need some information about their joint distribution. It follows that both $\{\ln(C/O)\}^2/\Delta t$ and $(Z_H + Z_L)/\Delta t$ provide unbiased estimators of the volatility parameter σ^2 and indeed the latter is independent of the former.

These estimators are areas illustrated in Figure 7.6. Consider the plot corresponding to time t. The vertical scale is logarithmic so that logs are plotted. This plot is constructed using an arbitrarily chosen angle θ from the four values (O, C, H, L) using two lines ℓ_1, ℓ_2 through the point $(t, \frac{1}{2}(\ \ln(O) + \ \ln(C)))$ with

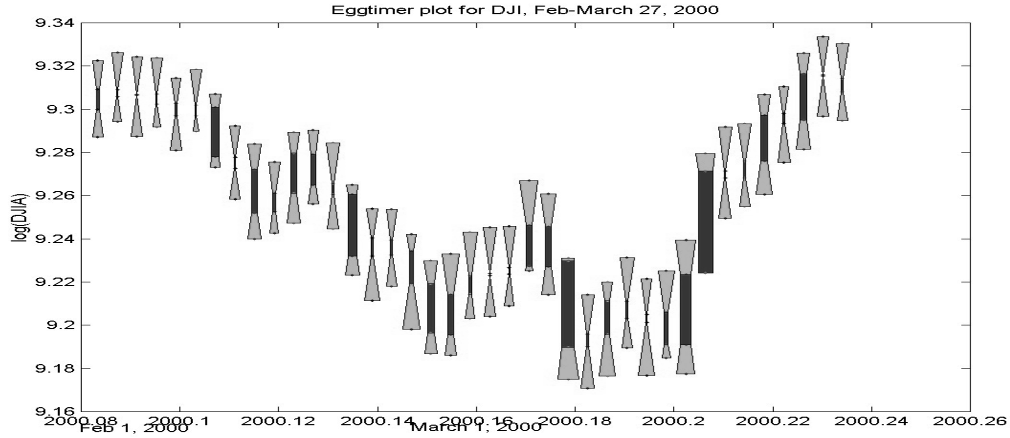

Figure 7.7: Eggtimer Plot for the Dow Jones Average, Feb 1-March 27, 2000.

slopes $\pm tan(\theta)$. Horizontal lines are drawn at the ordinate values $\ln(H)$, $\ln(L)$, $\ln(O)$, $\ln(C)$ and using the points where $\ln(O)$and $\log(C)$ strike the two lines as corners, a rectangle is constructed. The area of this rectangle $tan(\theta)(\ \ln(C/O))^2$ is an unbiased estimator of $tan(\theta)\sigma^2\Delta t$ provided the Brownian motion has no drift. The second region consists of "wings" generated by the four points at which the horizontal line at $\ln(H)$, $\ln(L)$ strike the lines ℓ_1, ℓ_2. The total area of this region (both wings) is $tan(\theta)(Z_L+Z_H)$ which is another unbiased estimator of $tan(\theta)\sigma^2T$ independent of the first, and also independent of whether or not the underlying Brownian motion has drift. By comparing these areas, we can detect abnormal changes in the volatility, or changes in the drift of the process that will increase the observed value of $(\ \ln(C/O))^2$ while leaving the second estimator unchanged. Because each estimator is based only on a single period, it is useful to provide as well a plot indicating whether there is a persistent change in either or both of the two estimators of volatility. Related estimators have been suggested in the literature.

We also show empirically the effectiveness of incorporating the high low close information in a measure of volatility. For example, the plot below gives the eggtimer plot for the Dow Jones Industrial Index for the months of February and March 2000. The vertical scale is logarithmic since the Black Scholes model is such that the logarithm of the index is Brownian motion. A preponderance of black rectangles shows periods when the drift dominates, whereas where the grey tails are larger, the volatility is evidenced more by large values of the high or small values of the low, compared to the daily change.

The 21-day rolling sum of the areas of the regions, either grey or black, is graphed in Figure 7.8 and provides two measures of volatility. The rolling sum of the grey areas is called the "intraday volatility" and that of the black

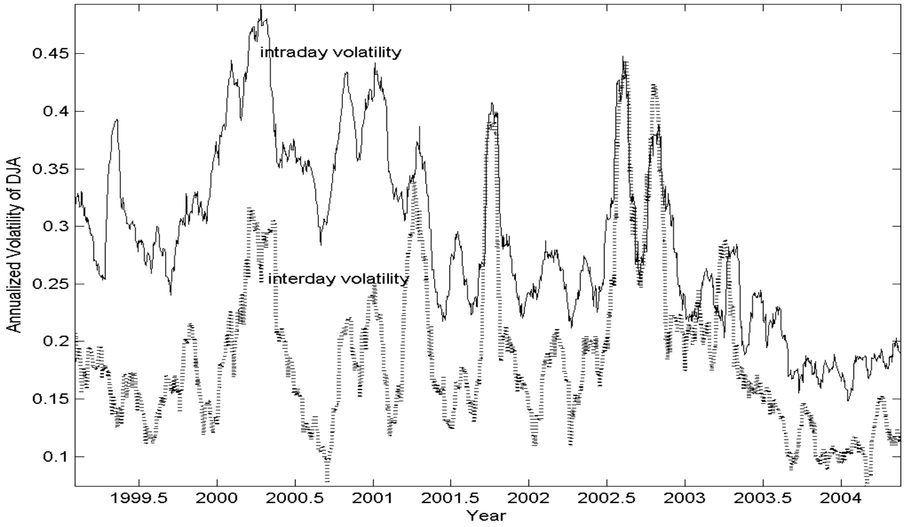

Figure 7.8: 21-day average volatility estimates for the Dow Jones Average, 1999-May 20, 2004

rectangles, the interday volatility. In the absence of substantial drift, both measure the same theoretical quantity but they differ in this case.

This difference is equally evident from Figure 7.9, the plot of the cumulative variance for the same period of time. Of course in this plot, it is the slope not the level which indicates the variance.

Consistent differences between the intra-day and the inter-day volatility or variances would be easy to explain if the situation were reversed and the inter-day volatility were greater, because one could argue that the inter-day measure contains a component due to the drift of the process and over this period there was a significant drift. A larger intraday volatility is more difficult to explain unless it is a failure of the Black-Scholes model. In this case, one might expect a similar behaviour in another market. If we generate a similar plot over the identical period of time for the NASDAQ index (Figure 7.10) we find that the comparison is reversed. This, of course, could be explained by the greater drift of the technology dependent NASDAQ (relative to its volatility) compared to the relatively traditional market of the Dow Jones but other indices we compared were more like the NASDAQ here than like the DJA.

There is no doubt that this difference reflects a greater intraday range for the DJA than other indices. In fact if we plot the cumulative value of the range of the index divided by the close $(H-L)/C$ as in Figure 7.11, it confirms that the daily range as measured by this ratio is consistently smaller for the NASDAQ

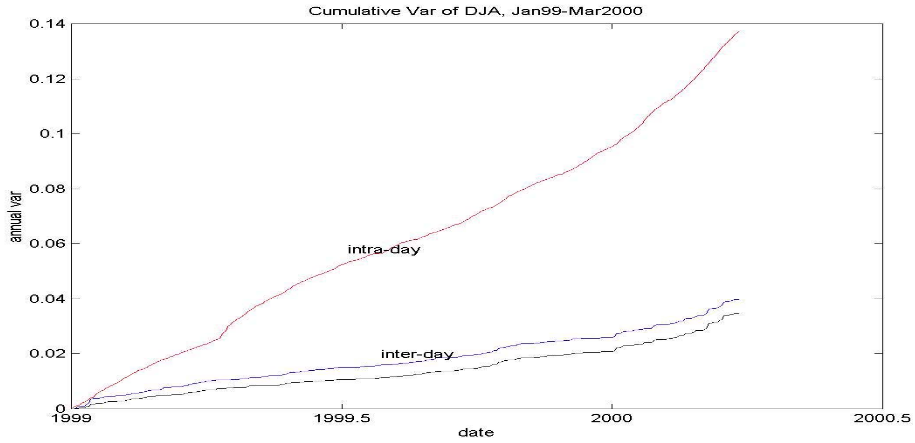

Figure 7.9: Cumulative Variances of the Dow Jones Average

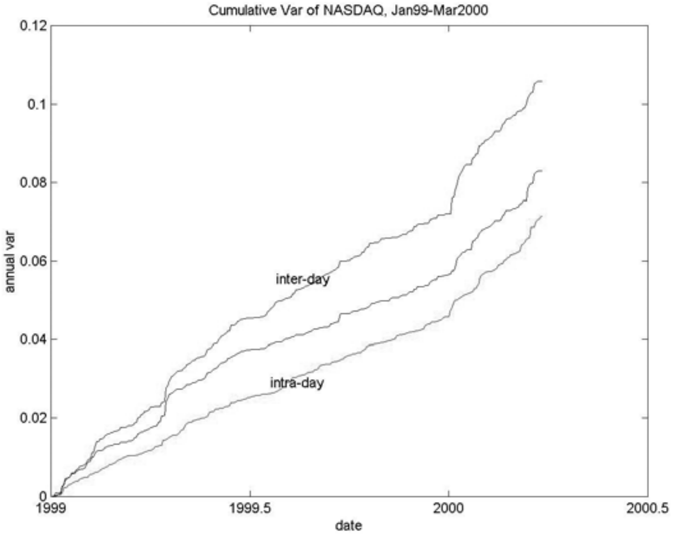

Figure 7.10: Cumulative Variances for the NASDAQ index

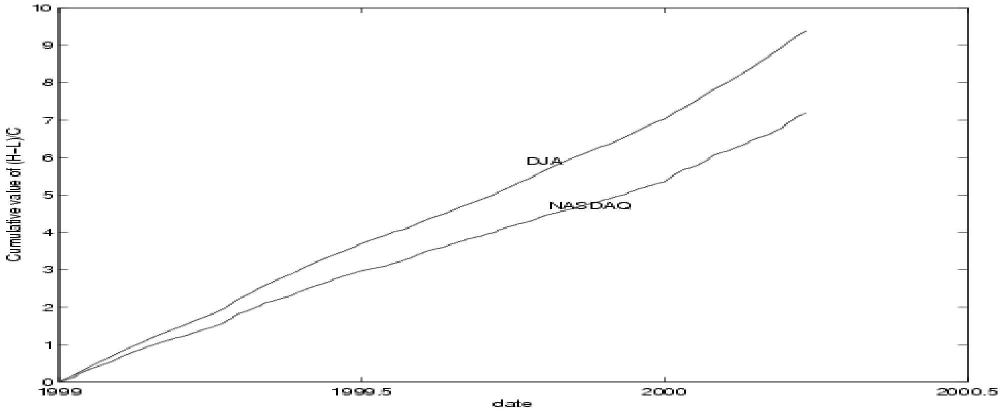

Figure 7.11: Comparison of the cumulative variance of the Dow Jones and the NASDAQ, Jan 1999-July 2000

than for the Dow Jones for this period.

Although high, low, open, close data is commonly available for many financial time series, the quality of the recording is often doubtful. When we used older data from the Toronto Stock Exchange, there were a number of days in which the high or low were so far from open and close to be explicable only as a recording error (often the difference was almost exactly $10). When the data on highs and lows is accurate, there is substantial improvement in efficiency and additional information available by using it. But there is no guarantee that published data is correct, particularly old data. A similar observation on NYSE data is made by Wiggins (1991); "*In terms of the CUPV data base itself, there appear to be a number of cases where the recorded high or low prices are significantly out of line relative to adjacent closing prices*".

7.7 Estimating Hedge ratios and Correlation Coefficients

The correlation coefficient between two asset prices is important not only because it indicates a degree of dependence between the two assets but because it is a required component for many practical investment decisions such as optimal portfolio selection, risk management, and hedging. There is no perfect hedge (except apparently in a Japanese garden) and so in practice we are required to use one asset to hedge another (hopefully highly correlated) asset. These may include derivatives or similar investments (for example bonds on the same or related underlying) which react similarly to market conditions.

Suppose we wish to hedge one investment, say in stock 2 using another, stock 1. As before, assume that we have data on the high, low, open and close of both stocks over non-overlapping time intervals $(t_i, t_{i+\Delta_i})$, $i = 1, 2, ..., n$. Denote the high, low, open and close for stock j on time interval i by $(H_{ij}, L_{ij}, O_{ij}, C_{ij}), i = 1, .., n$, $j = 1, 2$. Again we will denote the returns by

$$R_{ij} = \ln(C_{ij}/O_{ij}).$$

Then under the Geometric Brownian motion assumption, the return vector (R_{i1}, R_{i2}) has a bivariate normal distribution with variances $\sigma_1^2\Delta_i, \sigma_2^2\Delta_i$ and correlation coefficient ρ.

If we assume that we are able to rehedge our portfolio at the beginning of each time interval, and at the beginning of interval i we are long 1 unit worth of stock 2 and short h_i units worth of stock 1, then our total return at the end of the i'th period is $R_{i2} - h_i R_{i1}$. The optimal hedge ratio is the value of h_i which minimizes the variance of $R_{i2} - h_i R_{i1}$ and this is given by

$$h_i = \frac{cov(R_{i2}, R_{i1})}{var(R_{i1})} = \rho\frac{\sigma_2}{\sigma_1}.$$

Note that h_i is a constant independent of time, at least over periods when the stock volatilities and correlation coefficients remain unchanged. While implied volatilities σ_1 ,σ_2 may be obtained from derivative prices for each of these assets, the correlation parameter ρ is unknown and, unless there is a traded option such as a spread option whose value depends specifically on this correlation, ρ needs to be estimated from historical data. The simplest estimator of the correlation coefficient is the sample covariance of the returns,

$$\hat{\rho}_C = \widehat{cor}(R_{i2}, R_{i1}) = \frac{\sum_i (R_{i2} - \overline{R}_2)(R_{i1} - \overline{R}_1)}{\sqrt{\sum_i (R_{i2} - \overline{R}_2)\sum_i (R_{i1} - \overline{R}_1)}} \tag{7.59}$$

where $\widehat{cor}$ denotes the sample correlation coefficient. By a common argument (see for example Anderson (1958, Theorem 4.2.6) this has asymptotic variance

$$\frac{1}{n}(1-\rho^2)^2.$$

Here, as in McLeish(2004), we will consider using historical data for $(H_{ij}, L_{ij}, O_{ij}, C_{ij}), i = 1, .., n$, $j = 1, 2$ to estimate ρ. In the case of two or more correlated geometric Brownian motion processes, the joint distributions of highs, lows and closing values is unknown, and so we will need to revert to a simpler alternative than maximum likelihood estimation. We have seen that in the Black-Scholes model, the statistics

$$\begin{aligned}
Z_{Hi1} &= \ln(H_{i1}/O_{i1})\ln(H_{i1}/C_{i1}) \\
Z_{Hi2} &= \ln(H_{i2}/O_{i2})\ln(H_{i2}/C_{i2}) \\
Z_{Li1} &= \ln(L_{i1}/O_{i1})\ln(L_{i1}/C_{i1}) \\
Z_{Li2} &= \ln(L_{i2}/O_{i2})\ln(L_{i2}/C_{i2})
\end{aligned}$$

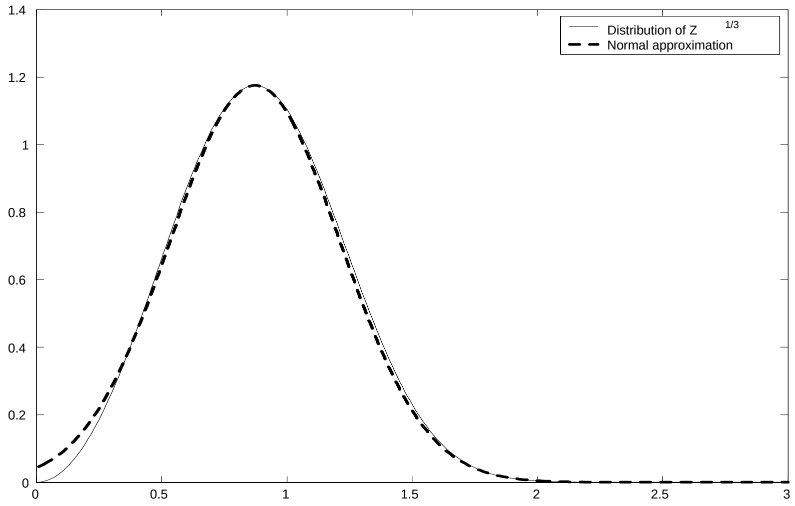

Figure 7.12: The Normal approximation (dashed line) to the distribution of $Z^{1/3}$ where Z is exponential.

all have marginal exponential distributions and each is independent of the open and close for the i'th interval.

Suppose we transform each of the above exponential random variables with some function g and assume that we can determine the correlation coefficients $cor(g(Z_{H1}), g(Z_{H2})) = b(\rho)$ as a function of ρ. For simplicity assume that we have subtracted the mean and divided by the standard deviation to provide a function g such that $E\{g(Z_{H1})\} = 0, var\{g(Z_{H1})\} = 1$. There are various possibilities for the transformation g, the simplest being a standardized power $g(Z_{H1}^p) = (Z_{H1}^p - E(Z_{H1}^p))/\sqrt{var(Z_{H1}^p)}$ for some suitable value of $p > 0$. A transformation of the gamma distributions in general and the exponential distribution in particular that make them very nearly normal is the cube root transformation ($p = 1/3$) (see Sprott, 2000, Chapter 9, Appendix).

For an exponential random variable Z_H there is only a slight difference between the graph of the probability density function of

$$\frac{Z_H^{1/3} - E(Z_H^{1/3})}{\sqrt{var(Z_H^{1/3})}} \tag{7.60}$$

and the graph of the standard normal probability density function, this difference lying in the left tail (see Figure 7.12), the region in which we are least interested in the maximum since it is very close to the open or close. We replaced the theoretical mean and variance in (7.60) by the sample mean and variance and so the function used ultimately was

$$g(Z_{Hi}) = \frac{Z_{Hi}^{1/3} - \overline{Z_H^{1/3}}}{\sqrt{\widehat{var}(Z_H^{1/3})}}.$$

Although it is theoretically possible to obtain the function $b(\rho)$, the exact expression requires the evaluation of Bessel functions. For simplicity we fit polynomials of varying degrees. For example the polynomial of degree 9

$$b(\rho) \approx .4822\rho^9 + .2102\rho^8 - .9629\rho^7 - .3104\rho^6 + .7006\rho^5 + .1887\rho^4 - .1288\rho^3 + .1822\rho^2 + .6385\rho + .0008$$

is accurate to within 1%.

Inverting this approximation to $b(\rho)$ provides a tractable estimator of the correlation between stocks based only on the correlation between the marginally exponential statistics Z_{Hi}, Z_{Li}. This estimator is

$$\hat{\rho}_{HL} = b^{-1}(\frac{1}{2}(\widehat{cor}(Z_{H1}^{1/3}, Z_{H2}^{1/3}) + \widehat{cor}(Z_{L1}^{1/3}, Z_{L2}^{1/3}))) \tag{7.14}$$

A similar estimator obtains as well from the cross terms since $cor_\rho(Z_{H1}^{1/3}, Z_{H2}^{1/3}) = b(-\rho)$.

$$\hat{\rho}_2 = -b^{-1}(\frac{1}{2}(\widehat{cor}(Z_{H1}^{1/3}, Z_{L2}^{1/3}) + \widehat{cor}(Z_{L1}^{1/3}, Z_{H2}^{1/3}))) \tag{7.15}$$

where again $\widehat{cor}$ denotes the sample correlation, but this estimator is very inefficient and adds little to our knowledge of the correlation. We will not consider it further. Since all that is required is the inverse of the function b, we can use a simple polynomial approximation to the inverse without sacrificing much precision, for example

$$b^{-1}(c) \approx 0.1567c^5 - 0.5202c^4 + 0.6393c^3 - 0.8695c^2 + 1.5941c.$$

The contenders for reasonable estimators are $\hat{\rho}_{HL}, \hat{\rho}_C$, or some combination such as an average, the simplest being a straight average of these two

$$\hat{\rho}_{AV} = \frac{1}{2}\hat{\rho}_C + \frac{1}{2}\hat{\rho}_{HL}.$$

It remains to see how these estimators perform in practice both on real and simulated data. First we simulated one year's worth of daily observations on two stocks, simulating the High, Low, Open, and Close for each day. There is considerable benefit to using both of the estimators $\hat{\rho}_C$ and $\hat{\rho}_{HL}$ at least for simulated data. For example in Figure 7.13 we graph the two variances $var(\hat{\rho}_C)$ and $var(\hat{\rho}_{AV})$ for various (positive) values of ρ. Notice that the estimator $\hat{\rho}_{AV}$ which combines information from all of the high, low, open, close has variance roughly one half of that of $\hat{\rho}_C$ which uses only the values of the open and the close.

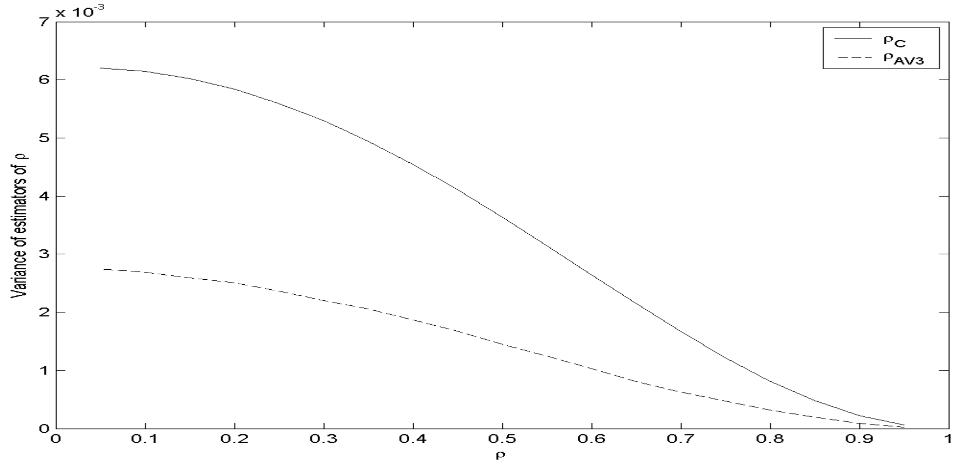

Figure 7.13: Variance of Estimators $\hat{\rho}_C$ and $\hat{\rho}_{AV}$ as a function of ρ.

We also use these methods to investigate the correlation between two processes. To this end we downloaded data for both the S&P500 index and Nortel (NT) over the period June 17, 1998-June 16, 2003. This was a period when Nortel rose and fell sharply showing increasing volatility. In Figure 7.14, we plot the two measures of the correlation $\hat{\rho}_{AV}$ and $\hat{\rho}_C$ between the returns for these two series in moving windows of 63 days (approximately 3 months) over this period. Note that the estimator $\hat{\rho}_{AV}$, indicated by a dotted line, tends to give lower correlation than $\hat{\rho}_C$ since it includes a component due to the movements within days and these have lower correlation than the the movements between days.

7.8 Problems

1. Suppose Y_1 and Y_2 are independent normal random variables with possibly different variances σ_1^2, σ_2^2 and expected values $E(Y_i) = \mu_i, i = 1, 2.$ Show that the conditional distribution of Y_1 given $Y_1 + Y_2 = S$ is Normal with mean
$$\mu_1 + w(S - \mu_1 - \mu_2)$$
and variance
$$\sigma_1^2(1 - w)$$
where
$$w = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}.$$

2. Consider data generated from the mixture density
$$pN(\theta_1, \sigma_1^2) + (1 - p)N(\theta_2, \sigma_2^2)$$

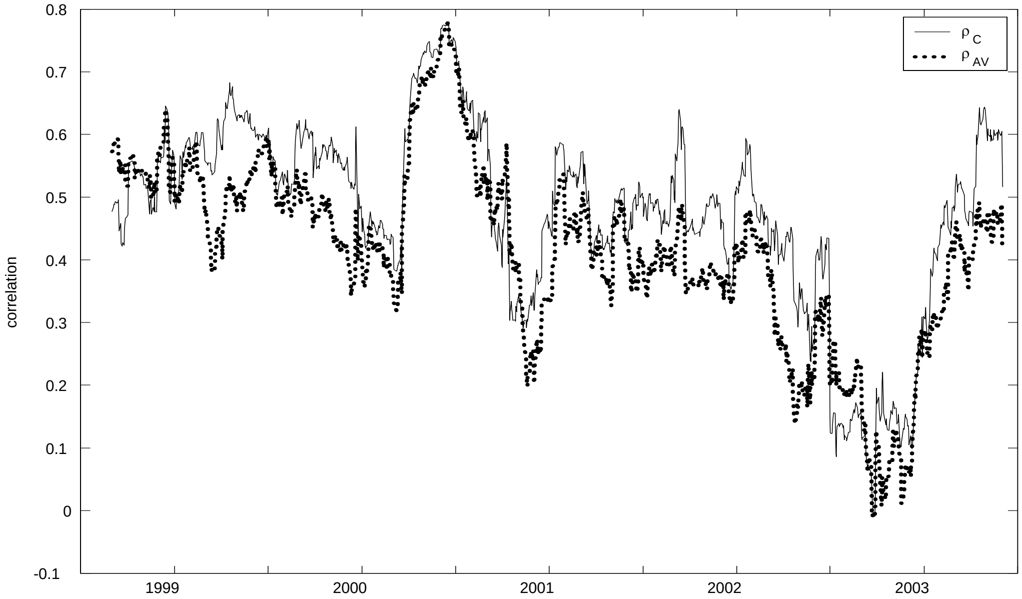

Figure 7.14: Two Measures of Correlation between Nortel (NT) and the S&P500 Index

where $\theta_1 < \theta_2$ (so that the parameters are identifiable). Write a program which uses the EM algorithm to estimate the five parameters in this model and test it on simulated data for sample sizes $n = 100, 1000, 2000$. Comment on bias and variance of the estimators.

3. Consider the following algorithm: the random variables ε denote standard normal random variables, independently generated at every occurrence:

    (a) Generate X_0 from an arbitrary distribution. Put $Y_0 = \rho X_0 + \sqrt{1-\rho^2}\varepsilon$

    (b) Repeat for $n = 1, 2, ...10, 000$

        i. Define $X_n = \rho Y_{n-1} + \sqrt{1-\rho^2}\varepsilon$

        ii. Define $Y_n = \rho X_n + \sqrt{1-\rho^2}\varepsilon$

    (c) Output $(X_j, Y_j), j = 5000, ...10000$
    Plot the points $(X_j, Y_j), j = 5000, ...10000$ and explain what this algorithm is designed to provide.

4. Suppose two interest rate derivatives (F and G) have price processes depending on the spot interest rate $r(t)$

$$\exp\{f(t, r(t))) \text{ and } \exp\{g(t, r(t)\} \text{ respectively,}$$

where, under the risk-neutral distribution,

$$df = \alpha_f(t)dt + \sigma_f(t)dW_t$$
$$dg = \alpha_g(t)dt + \sigma_g(t)dW_t.$$

In this case both derivatives are driven by the same Brownian motion process W_t which drives the interest rates $r(t)$. Show that if we maintain a proportion of our investment

$$\pi_t = \frac{\sigma_g(t)}{\sigma_g(t) - \sigma_f(t)}$$

(a negative value corresponds to short selling) in derivative F and the remainder $1 - \pi_t$ in derivative G, then our investment is risk-free and has value V_t satisfying

$$dV_t = V_t\{\pi_t \alpha_f(t) + (1 - \pi_t)\alpha_g(t)\}dt.$$

Therefore

$$\pi_t \alpha_f(t) + (1 - \pi_t)\alpha_g(t) = r(t)$$

and this implies a relationship between the drift and diffusion terms:

$$\frac{\alpha_g(t) - r(t)}{\sigma_g(t)} = \frac{\alpha_g(t) - r(t)}{\sigma_g(t)} = \lambda(t), \text{ say,}$$

with $\lambda(t)$ independent of the particular derivative. In other words all interest rate derivatives can be expressed in the form

$$df = [r(t) + \sigma_f(t)\lambda(t)]dt + \sigma_f(t)dW_t.$$

Assume you have available observations on the price of two interest rate derivatives taken daily over a period of 100 days as well as current interest rates $r(t)$. Assume the diffusion coefficients $\sigma_f(t), \sigma_g(t)$ do not depend on time and $\lambda(t)$ is a linear function of t. Explain how you would calibrate the parameters in this model to market data. Simulate data using constant values for $\sigma_f < \sigma_g$ and a constant value for $\lambda(t)$ and compare your estimates with the true values.

5. Assume a diffusion model for interest rates with time-varying coefficient;

$$dr_t = a(r_t, t)dt + \sigma(r_t, t)dW_t.$$

Consider a 0-coupon bond which, if invested today at time t returns 1 dollar at time T. If the current short rate is r_t, the value of this bond can be written as a function

$$f(r_t, t) = E_Q[exp\{-\int_t^T r_s ds\}]$$

where E_Q denotes expectation under the risk-neutral measure. The *yield curve* describes the current expectations for average interest rates;

$$\text{Yield}(T-t) = -\frac{log(f(r_t,t))}{T-t}$$

The more common models such as the Vasicek, the CIR and the Merton models for interest rate structure are such that the yield curve is *affine* or a linear function of the interest rate, i.e. $f(r,t) = exp\{c(T-t)+d(T-t)r\}$ for some functions $c(.), d(.)$. Generally this linearity occurs provided that both the drift term and the square of the diffusion coefficient $\sigma^2(x,t)$are linear in r. Suggest a graphicsl method for calibrating the parameters $c(T-t), d(T-t)$ to market data if we are provided with the prices of zero coupon bonds with a variety of maturities T at a number of time points $t_1 < t_2 < ...t_n$.