VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

ROBUST VOLATILITY ESTIMATION AND ANALYSIS OF THE LEVERAGE EFFECT

John Randal

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of

Doctor of Philosophy

For Jo and Kaitlyn

Abstract

Using volatility estimation as the underlying commonality, this thesis traverses the statistical problem of robust estimation of scale, through to the financial problem of valuing call options over stock.

We use a large simulation study of robust scale estimators to benchmark a non-parametric volatility estimation procedure, which not only uses techniques which are particularly suited to observed financial returns, but also addresses the problem of bias in any robust volatility estimation procedure.

Existing option pricing models are discussed with careful study of the assumed volatility and elasticity of volatility with respect to stock price relationships for each of these models. An option pricing formula is derived which extends existing methods, and provides a closed form solution which can be readily computed. Preliminary analysis of real price data suggests this model is able to explain observed leverage phenomena.

Acknowledgments

Many events conspired against the completion of this thesis, not least of which was my inability to stay in the saddle of my mountain bike.

I am particularly grateful to my supervisors, Peter Thomson and Martin Lally, for their invaluable guidance throughout this project. In particular, I would like to thank Peter for introducing the EM algorithm, and showing how we could use it to our advantage in both the simulations of Chapter 2 and the volatility estimation of Chapter 3, and Martin, for suggesting the combination of Rubinstein's (1983) model with that of Geske (1979), resulting in the pricing formula of Theorem 4.7, and also for his patience when I did not appear to be doing much finance at all. Thanks also to Robert Davies for helpful suggestions for the simulation of Chapter 2, and Clive Granger for a helpful discussion about volatility estimation.

Special thanks to all those involved in my upbringing and education to date, especially Mum and Dad.

Lastly, I must thank my family, Jo and Kaitlyn, for their love and support.

Contents

List of Tables

List of Figures

Chapter 1

Preliminaries

This thesis was motivated by an analysis of the constant elasticity of variance (CEV) option pricing model (Randal 1998). Empirical analysis therein focused heavily on the relationship between the log volatility series for a stock and the log stock price itself. This analysis highlighted two features of financial returns: the returns are heavy-tailed and have evolving volatility, as acknowledged in the volatility estimation literature. Surprisingly, we also found that this evolving volatility is difficult to estimate robustly, i.e. so that the estimates are unaffected by the underlying distribution of the returns. We also found that the relationships between volatility and price were not always consistent with basic financial theory. This thesis addresses those concerns.

Readers of this thesis are assumed to have a reasonable knowledge of statistical techniques, and also a knowledge of finance, and in particular, the area of option pricing. We begin wholly in one camp, and end up almost entirely in the other.

1.1 Structure

This thesis was originally meant to be a story of volatility and leverage. The only contender for a constant theme throughout is volatility: a measure of the variability of financial returns. We begin with the more general problem of estimating scale robustly. Following this, we focus directly on obtaining a robust volatility estimator, and finally we examine the underlying firm structure and posit a model which has the ability to explain a range of relationships between volatility and price level. The volatility estimator developed earlier is used to estimate such relationships, and thus to appraise the usefulness of the proposed model.

Replication of the Lax (1985) study on robust scale estimation in small samples was meant to be a very minor part of the work undertaken; however it became clear that a more extensive description of the work was in order. At the risk of it dominating this thesis, the majority of that work is given in Chapter 2 and supplementary material in Appendices B and C. In reading these sections, it may be useful to bear in mind that the goal throughout that work was to identify estimators which could usefully be used to estimate the variability of financial returns. Featuring in the simulations were estimators specifically designed to be used for estimation of volatility of financial price processes, and these are based on the t-distribution, which has been found to approximate the estimated distributions observed in practice. We find that these estimators are not only excellent for the purpose for which they were designed, but also good more generally.

The volatility estimator described in Chapter 3 appears to be successful. In the final section of this introductory chapter, we demonstrate both the difficulty in obtaining a robust time-series scale estimate, and also the importance of successfully doing so. The estimator we propose is based on the t-distribution with $\nu = 5$ degrees of freedom, since this distribution appears to be intermediate among the distributions observed in practice. The estimator is tested using simulation and a known evolving volatility process, for a variety of distributions. It performs very well in these simulations, and is benchmarked against the traditional non-robust volatility estimation procedure (based on a moving standard deviation), and also the best-performing estimator identified in the simulations of Chapter 2. The estimator is also applied to real data, and further properties of the volatility estimates are identified.

Having secured a robust and reliable volatility estimator, we move in Chapter 4 to analyse the relationship volatility and price have, both in theory and in practice. Several theoretical models are included in this study, including the fundamental model of a firm with risky debt, and the stock price models assumed for the Black-Scholes, CEV, compound and displaced diffusion option pricing models. A new option pricing model is derived, and this both combines and extends the compound and displaced diffusion models. Significantly, we obtain a closed form solution for an option price under this model, which can be readily computed. Further, we find this model has the additional flexibility to model a variety of relationships between volatility and price. Analysis for Telecom New Zealand data shows the model is broadly consistent with what is observed.

Chapter 5 concludes with a summary of what has been found, and an indication of future research directions.

The remainder of this Chapter illustrates the difficulty of measuring scale of time series data robustly, in particular using the non-parametric smoother `loess`.

1.2 Using loess to estimate variability

The time series smoother `loess` (Cleveland, Grosse & Shyu 1992) can be used to provide a robust estimate of the level of a time series through time. It is implemented in the statistical software `R` (Ihaka & Gentleman 1996), and its details are discussed in Appendix A. As demonstrated in the appendix, `loess` can be used to provide a smooth level estimate that is not unduly affected by the occasional outlying values, nor by non-Gaussian data. This suggests that a natural way to estimate volatility using `loess` would be to smooth the squared returns obtained from financial asset price series.

It turns out that this is not such a good thing to do. The robust estimate is obtained from `loess` by specifying `family="symmetric"` in the function call, since for the robust estimate, `loess` merely assumes that the series we are smoothing is symmetric about the level we hope to estimate, rather than normally distributed about that level. Asset returns are typically symmetrically distributed, at least close to the mode of the distribution, and so it follows that the squared returns will be not at all symmetric. An example of this is shown in Figure 1.1 for a simulated geometric Brownian motion (GBM) series with $\mu = \frac{1}{2}\sigma^2$ and $\sigma = 1$. This process is consistent with log-normally distributed prices, and normally distributed returns, and is the process assumed by Black & Scholes (1973) in the derivation of their famous call option pricing formula. Under this choice of parameters, the daily returns are independent $\mathcal{N}(0, 1)$ variables. The left histogram in Figure 1.1 is the sample distribution of the returns for the simulated series, and these are fitted very well by the underlying normal distribution, as we would expect. The right histogram is of the squared returns. These have a chi-squared distribution with one degree of freedom, and this is superimposed. We see two dominant effects, all the returns less than one in absolute value are pushed toward zero, and the squared returns are highly asymmetric.

The effect of ignoring the requirement of a symmetric series is shown in Figure 1.2 for the simulated GBM series whose return distribution features in Figure 1.1. Even

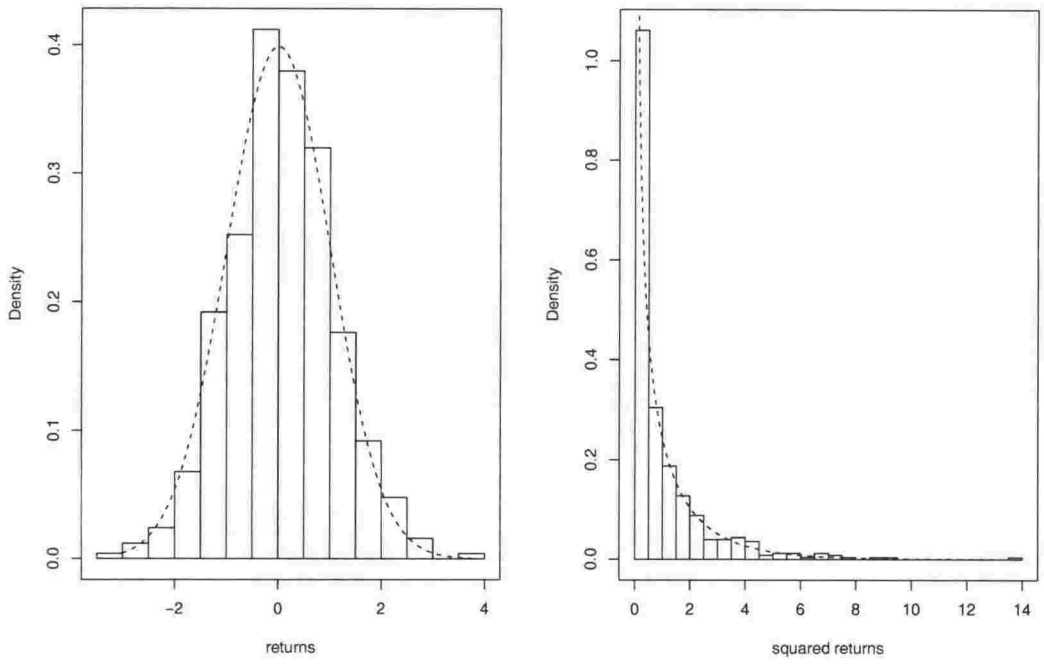

Figure 1.1. Distributions of returns and squared returns for a simulated geometric Brownian motion series 501 observations long, with $\mu = 0$ and $\sigma = 1$. The left histogram is for the returns and the standard normal density function is superimposed. The right histogram is for the squared returns and the χ^2_1 density function is superimposed.

though the underlying distribution of returns is Gaussian, and strictly speaking, robust estimation is not required, we see that the robust volatility estimate drastically under-estimates the true volatility function. This bias is caused by the asymmetric distribution of the squared returns, and the effect this has on the robustness weights used by `loess`. As described in Appendix A, the robustness weights are based on the previous iterate's residuals. These are divided by six times the median absolute deviation of the residuals from their median, and then inserted into the biweight function

$$B(u) = \begin{cases} (1-u^2)^2 & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

to obtain the robustness weights. The biweight function has a maximum at $u = 0$ of 1, and so the robustness weights are less than or equal to one.

The robustness weights for the squared returns on which Figure 1.2 is based, are plotted in Figure 1.3 against the residuals $(R_t^2 - \hat{\sigma}_t^2)$ from the robust fit (the largest residual 13.34 is omitted from the plot). The points trace out the biweight function, but the residuals are clearly not symmetrically distributed about zero. In particular, the residuals in the short left tail of the residual distribution all get relatively high robustness weight, with the minimum residual getting a robustness weight of 0.905.

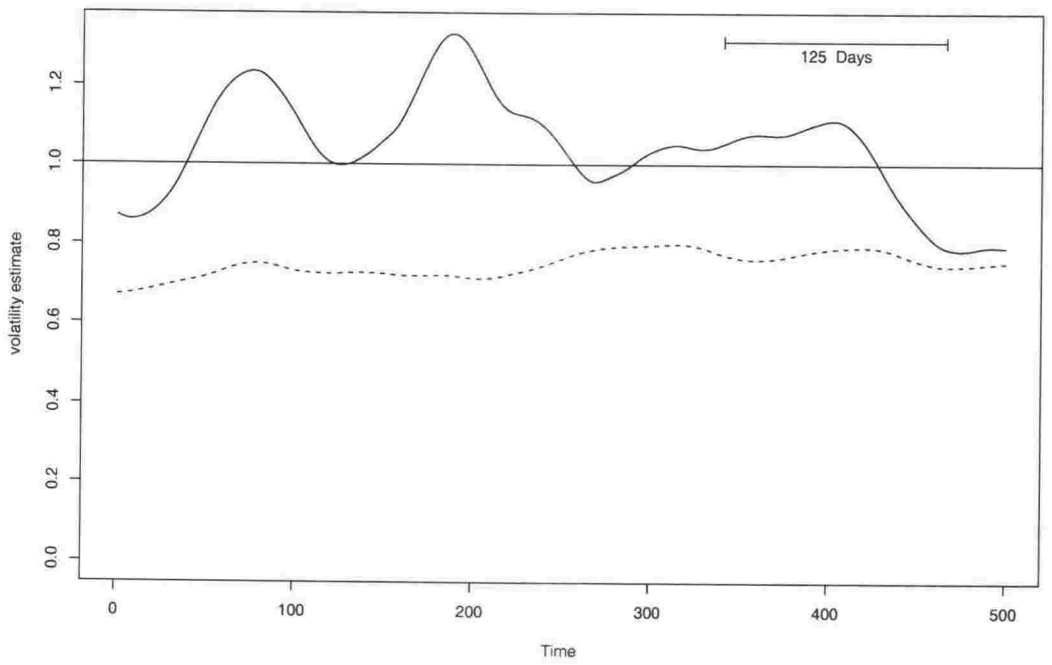

Figure 1.2. Volatility estimates using `loess` for the GBM series used for Figure 1.1. The solid estimate is from `loess` using standard smoothing techniques with no robustness properties. The dotted line is from `loess` with robust smoothing. In both cases the squared returns are smoothed with a window of 125 observations, and all estimates shown are based on a complete smoothing window. The true parameter is shown by the horizontal line.

In contrast, 25% of the observations get a robustness weight less than this, and all of these have positive residuals. Approximately 20% of the squared returns get robustness weight less than 0.8, and 6.6% of them get zero weight. Since the weights are strictly less than one, and the series we are smoothing is not symmetric (and so the weights do not offset either side of the centre), the volatility estimate is downwards biased as clearly demonstrated in Figure 1.2.

We conclude that because in general, squared returns will not be symmetrically distributed, as in the data analysed in this section, the robust smoother `loess` will not produce a robust estimate of volatility simply by smoothing the squared returns since it is based on a symmetric distribution about the level we are estimating. In the following two chapters, we examine alternative estimation techniques.

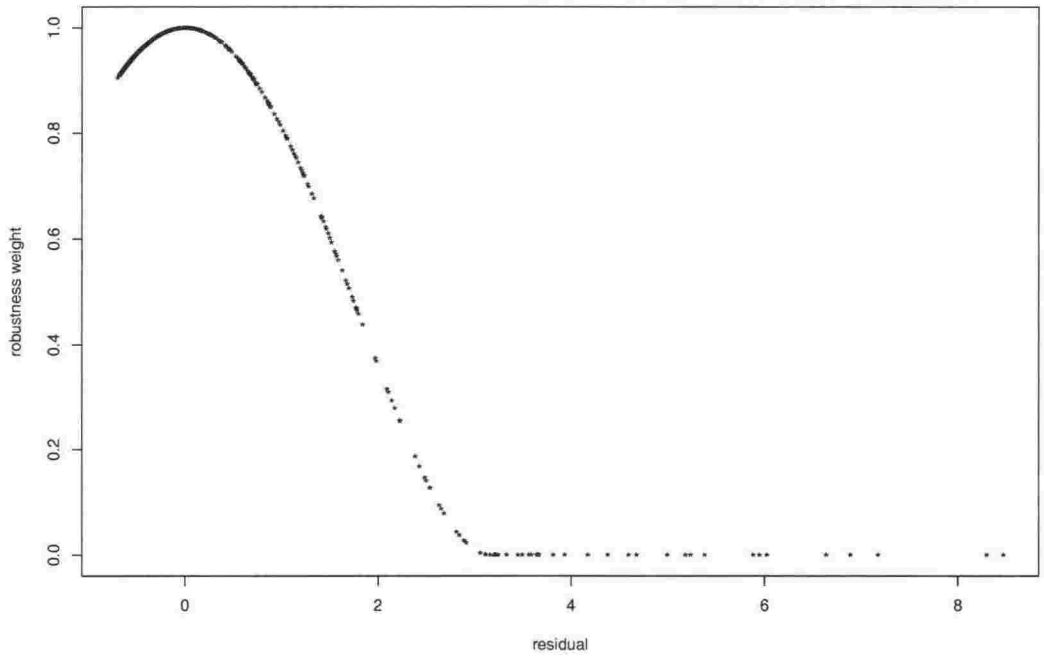

Figure 1.3. Robustness weights for the GBM returns shown plotted against the residual from the robust estimate of their variability. A single residual at 13.34 is omitted from the plot, and this has an associated robustness weight of zero.

Chapter 2

A reinvestigation of robust scale estimation in finite samples

"In analyzing data, we do not want to even attempt to represent its stochastic behaviour accurately; rather we wish to choose techniques that spare us this essentially impossible task" Morgenthaler & Tukey (1991), page 1.

With both the aim of measuring the variability of financial returns, and the above quote from Morgenthaler & Tukey (1991) in mind, we perform a simulation study of robust estimators of scale.

2.1 Introduction

This study examines robust scale estimation, using computer simulations which take advantage of recent improvements in computing technology. For this simple reason, simulations undertaken here dwarf those of Lax (1985) (hereafter referred to as Lax, or the Lax study), who performed a simulation study with identical purposes. The estimators and techniques considered here are motivated by the Lax study, and also by published studies using robust scale estimators in more recent times.

A robust estimator is called *resistant*, if it is largely unaffected by a small number of large changes to the data (i.e. by outliers) and by any number of small errors (e.g. rounding errors). Typically, we are more interested in resistance to outliers. In addition to possessing resistance properties, a robust estimator will be a suitable estimator for non-normal data (i.e. have high relative efficiency). The notion

of efficiency, and in particular for a scale estimator, will be clarified later, however at this stage it is sufficient to think of an "efficient" estimator having a mean squared error which is close to the minimum for a variety of situations (describing possible underlying distributions for data). Robust estimators will be particularly applicable for financial data, which often features the three situations we are protecting against: occasional rogue values, many small errors (induced by properties of financial markets such as discrete price intervals and discontinuous trading) and underlying non-normality.

The estimators considered are assessed by their minimum relative efficiency over Tukey's three corners: the standard normal distribution, the one-wild situation (also known as 1-wider), where $n-1$ of the observations in a sample of size n are standard normal and the remaining observation has 10 times the standard deviation of the others, and the slash distribution, an observation from which is obtained by dividing a standard normal random variable by an independent random variable distributed uniformly on the interval $[0, 1]$. These three sampling situations were considered by Tukey to reflect the three extreme cases of importance to robust statistics. All three theoretical distributions are symmetric: the normal has rapidly decaying tails; the one-wild allows the presence of a single outlying, but otherwise well behaved, value (in the upper or lower tail with equal probability); and the slash, with its infinite mean and variance, has very slowly decaying tails. In practice, most samples from the one-wild will be highly asymmetric, with the presence of the single outlier. An estimator which copes well in all three situations can suitably be used:

- when the data is well behaved;
- in the presence of occasional outliers;
- when the data is very heavy tailed;
- or some combination (see Yatrakos 1991)

and is thus highly useful, particularly when much data is being processed with little interaction by the analyst.

In the remainder of this introduction we present a discussion of some of the important considerations when conducting a simulation study of this sort, and a summary of the current state of the robust scale estimation literature. In the following sections

we describe the estimators and methodology of this simulation study, and in the final section of this chapter, the results are presented and discussed. Supplementary material is given in Appendices B and C.

2.1.1 Tukey's three corner distributions

Tukey's three corner distributions are intended to model extreme behaviour for data. An estimator which performs well in simulations for each of these distributions is likely to perform well for any data met in practice. The properties of the three distributions are discussed further in the following notes.

The normal distribution

In classical statistics, data are all too often assumed to be drawn from normal distributions. While this may be an appropriate description for some types of data, such samples are arguably more the exception than the rule. The normal distribution is described by Morgenthaler & Tukey (1991) as "unrealistically nice" (page 7), and they prefer to use the descriptor "Gaussian" so as not to infer normality (in the lay sense) on the situation. Students rote learn the 68-95-99 rule: 68% of normal data lie within one standard deviation of the mean, 95% within two, and 99% within three, and the majority of statistical theory is based on the hope that this is a fair description of the population from which the data are drawn. This is of course reasonable in the many situations that the central limit theorem applies, for example when the sample mean is used for inference about the population mean. However, even in this case, non-normal data can cause problems if the sample standard deviation must be used as an approximation to the population standard deviation. The normal distribution has two parameters: the mean μ and the variance σ^2, and is often denoted $\mathcal{N}(\mu, \sigma^2)$. The standard normal distribution has $\mu = 0$ and $\sigma^2 = 1$.

There is not a great deal of leeway if optimality is based on underlying normality. Lax describes an experiment by Tukey in which a sample from a contaminated normal distribution is examined.

Definition 2.1 (Contaminated normal random variable) *The contaminated normal random variable X, with parameters $0 < p < 1$ and $k > 1$ and denoted $CN(p; k)$, is standard normal with probability $1 - p$, and otherwise normally distributed with mean 0 and variance k^2.*

With appropriate scaling, the contaminated normal distribution can be used to generate an observation from the normal distribution with parameters μ and σ^2 with probability $1-p$, and an observation from the normal distribution with parameters μ and $(k\sigma)^2$ with probability p. In this thesis, where used, p and k will generally be specified, and μ and σ^2 considered unknown. Thus, the contaminated normal is a mixture of two normal distributions, and a sample from this distribution can be considered "contaminated" in the sense that any observations drawn from the second distribution have replaced observations from the "correct" distribution with the lower variance.

Tukey's experiment showed that when $k = 3$ and the mixing parameter p exceeded 0.18%, the sample standard deviation, optimal for normal data, becomes asymptotically less efficient than using the mean absolute deviation about the mean, which has an 87.6% asymptotic efficiency for uncontaminated normal data. This level of contamination represents one observation in approximately 556.

Although normally distributed samples might be rare in practice, the normal distribution represents a reference case, and hence it is included as one of the three corners. The remaining corners represent departure from this ideal.

The one-wild sample

The one-wild sample is not well known outside the robust literature, and it is the subject of the following definition.

Definition 2.2 (One-wild sample) *The one-wild sample consists of $n-1$ observations drawn independently from the normal distribution with mean μ and variance σ^2, and a single observation drawn independently from a normal distribution with mean μ and variance $100\sigma^2$. A one-wild sample with $\mu = 0$ and $\sigma^2 = 1$ is called a standard one-wild sample.*

Without loss of generality, analysis of one-wild samples in this thesis is almost exclusively restricted to standard one-wild samples.

A one-wild sample is similar to a sample from the mixture of two normals with mixing parameter $p = \frac{1}{n}$ and scale factor $k = 10$, i.e. the probability of any observation being from the $\mathcal{N}(\mu, 100\sigma^2)$ distribution is $\frac{1}{n}$, and the probability of any observation being $\mathcal{N}(\mu, \sigma^2)$ is $1 - \frac{1}{n}$. The difference between samples from the two cases

is that while the one-wild will have a single "wild" (or contaminated) observation, the number in the sample from the contaminating distribution is Binomial, with n trials, and probability $p = \frac{1}{n}$. Thus, in a sample of size $n = 20$ from the mixture, we would *expect* only a single outlier, but the actual number N has the distribution:

x	0	1	2	3	4	5	6	$\cdots$	20
$P(N = x)$	0.358	0.377	0.189	0.060	0.013	0.002	0.000	$\cdots$	0.000

where the probabilities are rounded to three decimal places. Use of the one-wild, rather than the mixture, allows us to focus on the resistance properties of the estimators, without the results being influenced by approximately one third of the samples with no outliers, and another quarter with at least two outliers.

Kafadar (1982) points out that unlike the mixture, a single observation from the one-wild sample is not drawn from a single distribution, and hence the sample is not a random sample which comprises independent and identically distributed observations. Despite this, its use is favoured over that of the mixture, because it presents a consistent challenge, rather than a stochastic one (see Cohen (1991) for discussion). This behaviour is confirmed in Figures 2.1 and 2.2.

Figure 2.1 is based on the log sample standard deviation estimates from 20000 independent samples from the mixture $\mathrm{CN}(\frac{1}{20}, 10)$. These statistics have a bimodal distribution, and this reflects the behaviour of the standard deviation for the samples with 0, 1, 2 or more "outliers". In particular, the sample distribution of the standard deviations from the uncontaminated standard normal distribution (representing 7240 out of the 20000 samples) is superimposed, and we see this has a nice unimodal shape. The distribution of the standard deviations from this group and the one-wild samples (representing 7515 out of the 20000 samples) takes on the bimodal form of the whole group due to the poor performance of the standard deviation for the second group. This effect is magnified as the remaining samples are included.

The actual sampling distribution of the log standard deviations from the 7515 one-wild samples is shown in Figure 2.2, and this has a unimodal distribution. Focus on the one-wild, rather than the mixture, examines the response of an estimator to a consistent challenge, rather than the variety of situations embodied in Figure 2.1.

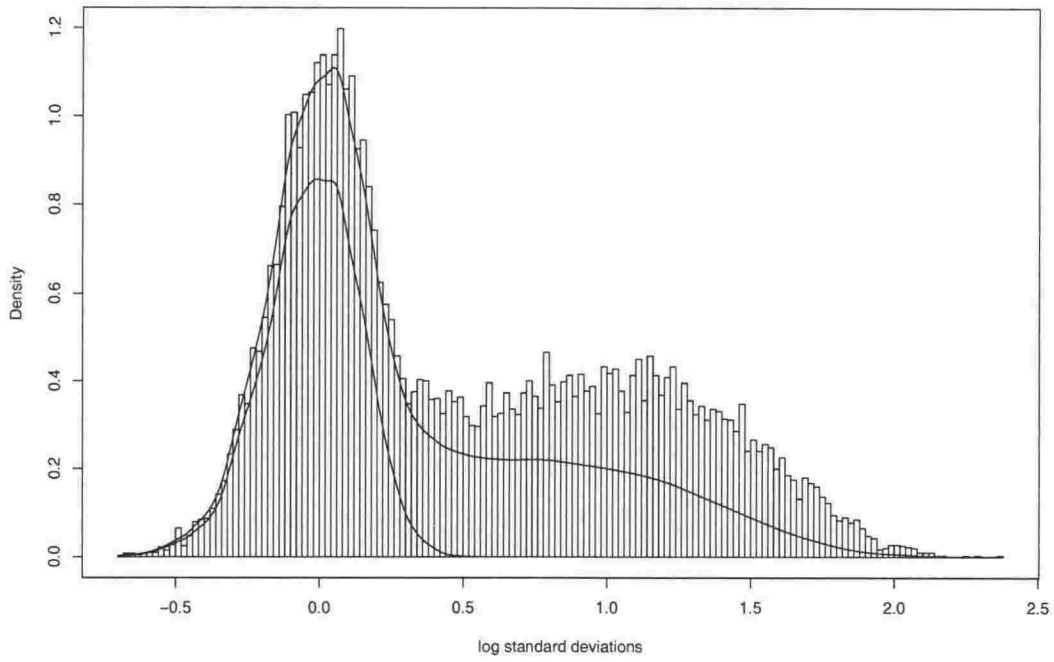

Figure 2.1. The distribution of log sample standard deviation for 20000 samples from the mixture $CN(\frac{1}{20}, 10)$. Superimposed are (scaled) estimated densities for the samples with 0 outliers, and for those samples with 0 or 1 outliers. The scaling is done so that each curve approximates the contribution in the histogram of the respective samples.

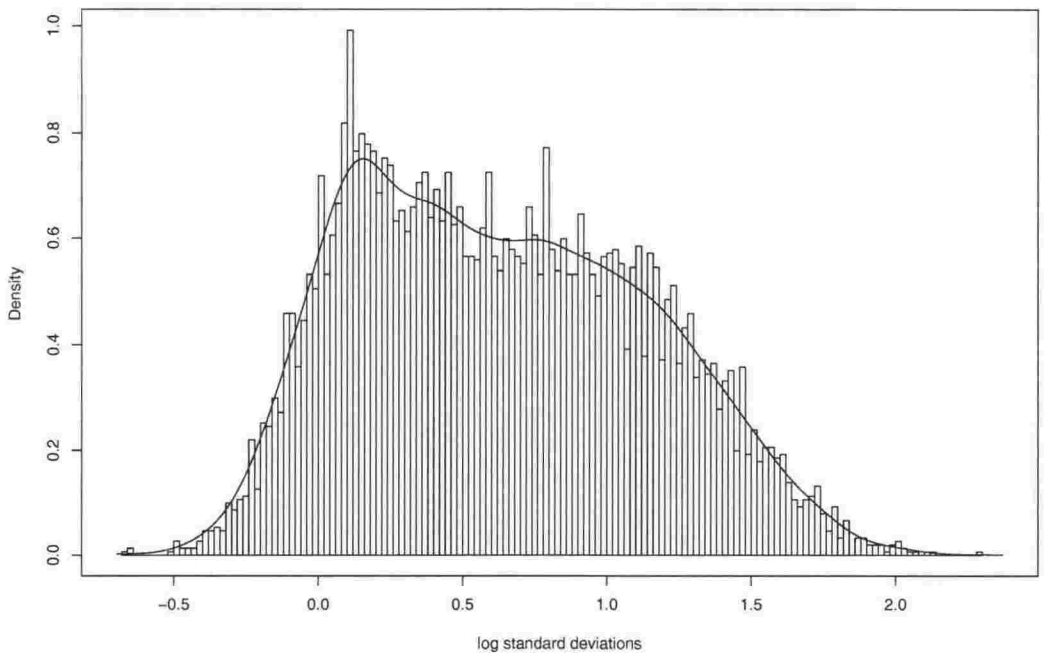

Figure 2.2. The distribution of log sample standard deviation for 7515 one-wild samples. These statistics are a subset of those shown in Figure 2.1, which were based on a random number of "wild" observations in each sample. Superimposed is the estimated density of these statistics.

The slash distribution

The slash distribution is well known in the robust statistics literature, but is less familiar generally than other long-tailed distributions such as the contaminated normal, Student's t-distribution, the double exponential, and the Cauchy. Like the Cauchy, the slash has no mean or variance due to its slowly decaying tails; however it is symmetric about its median and has a well defined scale parameter.

Definition 2.3 (Slash random variable) *The slash random variable X is defined as $X = \mu + \sigma\frac{Z}{U}$, where $\sigma > 0$, Z is a standard normal random variable, and U is an independently distributed uniform random variable on the interval $[0, 1]$. A slash random variable with $\mu = 0$ and $\sigma^2 = 1$ is called a standard slash random variable.*

Without loss of generality, analysis of the slash distribution in this thesis is almost exclusively restricted to the standard slash distribution.

The standard slash random variable has the density function

$$f_X(x) = \begin{cases} \frac{1}{x^2\sqrt{2\pi}}\left(1 - e^{-\frac{1}{2}x^2}\right) & x \neq 0 \\ \frac{1}{2\sqrt{2\pi}} & x = 0 \end{cases}$$

and the distribution function

$$F_X(x) = \begin{cases} \Phi(x) + \frac{1}{x}\left(\phi(x) - \frac{1}{\sqrt{2\pi}}\right) & x \neq 0 \\ \frac{1}{2} & x = 0 \end{cases}$$

where $\phi(x)$ and $\Phi(x)$ are the standard normal probability density and distribution functions respectively. The slash distribution can be compared to other, more familiar, long-tailed distributions by way of the tail weight index used in Rosenberger & Gasko (2000), who define the tail weight index $\tau(F)$ for the distribution F as

$$\tau(F) = \frac{F^{-1}(0.99) - F^{-1}(0.5)}{F^{-1}(0.75) - F^{-1}(0.5)} \times \frac{\Phi^{-1}(0.75) - \Phi^{-1}(0.5)}{\Phi^{-1}(0.99) - \Phi^{-1}(0.5)}. \tag{2.1}$$

This measures the ratio of the distances from the 99th and 75th percentiles to the median of F, rescaled by the same quantity for the normal distribution. Of course $\Phi^{-1}(0.5) = 0$ however these terms are left in (2.1) to emphasize the symmetry in the equation. We see $\tau(\Phi) = 1$ and distributions with longer tails than the normal's will have $\tau(F) > 1$. Rosenberger & Gasko (2000) provide a table of the tail weight

Distribution	$\tau(F)$	Kurtosis
Uniform	0·568	−1·2
Triangular	0·850	−0·6
Gaussian	1	0
t_{10}	1·145	1
CN($\frac{1}{20}$; 3)	1·204	0·471
Logistic	1·213	1·2
t_5	1·343	6
Double exponential	1·636	3
t_2	2·473	∞
CN($\frac{1}{20}$; 10)	3·429	5·490
Slash	7·866	∞
Cauchy (t_1)	9·226	∞

Table 2.1. Measures of tail length for some standard distributions. t_ν is the Student's t distribution with ν degrees of freedom, and CN($p; k$) is the contaminated normal of Definition 2.1. $\tau(F)$ is defined in (2.1), kurtosis in (2.2).

index for some well known distributions. This information is reproduced in Table 2.1 (correct to three, rather than two, decimal places), along with results for various Student's t-distributions.

Coefficients of kurtosis for these distributions are also given in the table. The coefficient of kurtosis is the fourth central standardised moment of the distribution of X, translated so that the normal distribution has zero kurtosis, i.e.

$$\kappa(X) = \frac{\mathrm{E}\{(X - \mathrm{E}(X))^4\}}{\mathrm{E}\{(X - \mathrm{E}(X))^2\}^2} - 3. \tag{2.2}$$

As is evident from the table, the slash distribution has very heavy tails, as does the contaminated normal CN($\frac{1}{20}$; 10). The latter distribution has implications for a one-wild sample which we would expect to have similar long-tailed behaviour. The infinite kurtosis shared by the Student's t-distributions with $\nu \leq 4$ and the slash distribution shows more dramatically than the tail weight index how long the tails of these distributions are. The contaminated normal maintains long yet well-behaved tails. We note that there is not a perfect rank-order correlation between the two measures. Also, while CN($\frac{1}{20}$; 3) and the logistic distributions have similar tail weight indices, their measures of kurtosis are quite different.

The slash distribution is an important one for the testing of robust estimators, since it is both easy to simulate, and it represents a worst case scenario for symmetrically distributed data. The Cauchy distribution provides an alternative, but the simplicity of the slash random variable makes it a more popular choice. In addition, and

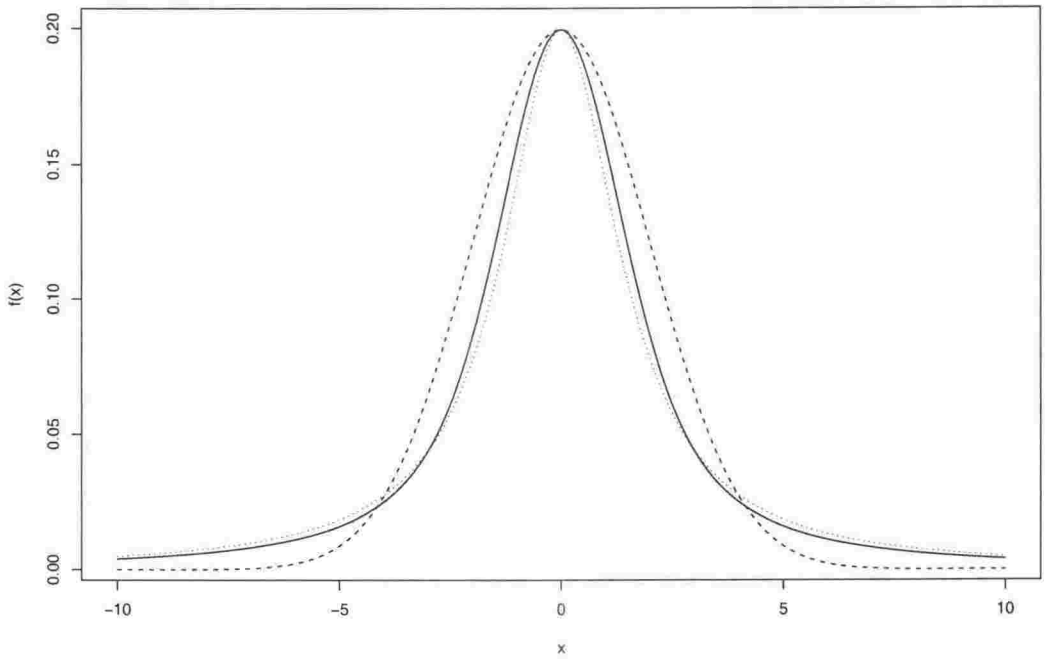

Figure 2.3. Probability density functions for the standard slash, Cauchy and normal random variables. The slash is given by the solid line, the Cauchy by the dotted line, and the normal by the dashed line. All density functions have mode at $x = 0$ with $f(0) = 1/(2\sqrt{2\pi})$. Thus the Cauchy has scale parameter $\sigma = 2\sqrt{2/\pi}$, and the normal has $\sigma = 2$.

perhaps of greater importance, the slash is more like the normal distribution for small x, as demonstrated by Rogers & Tukey (1972) and in Figure 2.3. In this plot we compare the density function of the standard slash, with the density function of the Cauchy with scale parameter $\sigma = 2\sqrt{2/\pi}$, and that of the normal with $\sigma = 2$. These scale parameters are chosen so that $f(0) = 1/(2\sqrt{2\pi})$ for all three distributions. It is evident that the slash and Cauchy densities are very similar, although the slash is closer to the normal around the mode.

The slash may well be too extreme a situation for most real data, but by tuning estimators to perform well for this distribution as well as for Gaussian data, and data featuring the occasional rogue value, we ensure high quality estimates regardless of the actual distribution of the data. This is the principle behind Tukey's triefficiency, discussed in the next section.

2.1.2 Triefficiency

An estimator's overall quality was assessed by Lax using the triefficiency promoted by Tukey.

Definition 2.4 (Triefficiency) *An estimator's triefficiency is the smallest of its efficiencies at each of normal, one-wild and slash samples of size n.*

The triefficiency is simply the minimum efficiency of the estimator over the three corners, and the "best" (triefficient) estimator will have the maximum triefficiency. We would expect the triefficiency of this estimator to be less than 100% since no single estimator will be optimal at all three corners. Should we have some data whose distribution is unknown, but not as "extreme" as one of the three corners, the triefficient estimator will give us a "good" estimate of scale for this data, regardless of the actual distribution.

To further enhance the importance of the triefficiency criterion, Yatrakos (1991) states that for any linear combination of the three corner distributions, any estimator will have efficiency at this distribution at least as great as its triefficiency. If we believe that the corners are indeed the extremes, then we can be confident in using the estimators that perform well in the simulations that are presented in Section 2.5. This is not to say that we would always wish to use the triefficient estimator. Clearly if we knew the data was Gaussian, we would certainly use the sample standard deviation over any of the robust estimators considered here, and if slash, we would definitely use a robust estimator (or indeed the ML estimator for the slash distribution). Thus, given knowledge of the actual distribution of the data, it is likely that we would not use the triefficient estimator, but an estimator particularly useful in that case. The results given in Section 2.5 will not only identify very good general purpose scale estimators, but also indicate which estimators are appropriate if prior knowledge of the distribution from which the data is drawn is known.

If efficiency is measured relative to a sub-optimal estimator for a single distribution, provided this is made clear, the efficiency measure is meaningful, e.g. if we know that estimator A is 85% efficient with respect to estimator B, this is useful information even if B is not optimal. This is not so clear if the triefficiency criterion is used, since if efficiency is relative to a sub-optimal estimator for a particular corner distribution, the efficiency in this case will be inflated, and the triefficiency may be too large.

Thus, there are a number of further considerations to make when using the triefficiency criterion. Perhaps the biggest flaw in use of the triefficiency criterion in the Lax study is its dependence on the estimators considered. Often in Lax's results (reproduced later in the text, in Table 2.3), an estimator's minimum efficiency is at

either the slash distribution (for non-robust estimators like the standard deviation) or at the normal distribution (for robust estimators designed specifically to mitigate the impact of long tails). The sample standard deviation is efficient for Gaussian data, and so over a large number of samples, it will have the minimum sample variance among estimators. Hence, relative efficiencies for Gaussian data based on the sampling variance of the standard deviation will be independent of the remaining estimators considered. In contrast, unless the minimum variance estimator for the slash and one-wild distributions are considered in the simulation, efficiencies for these distributions will be too high, and selection on the basis of triefficiency will be flawed.

We conduct a small experiment with Lax's results in order to examine the impact of possibly overstated efficiencies in the one-wild and slash cases. If the minimum variances used by Lax to compute the efficiencies are correct, 13 of the estimators have minimum efficiency at the normal distribution, one at the one-wild and the remaining three at the slash distribution. If the minimum variance estimator used by Lax for either the one-wild or slash corners is not the true minimum variance estimator, then it will have efficiency less than 100% relative to the true minimum variance estimator. We investigate the effect of this possibility on the triefficiency in Table 2.2. Specifically we count the number of times that each corner yields the minimum efficiency (the triefficiency) out of the 17 estimators given by Lax, for various combinations of one-wild and slash inefficiency. In particular, if the one-wild minimum variance is overstated such that the minimum variance estimator is only 80% efficient, we see the one-wild dominates the triefficiencies, even if the slash minimum variance estimator is itself only 80% efficient.

Use of the EM algorithm (Dempster, Laird & Rubin 1977), which yields the maximum likelihood estimates, allows the proper efficiencies to be calculated, and hence one of the shortcomings of Lax's study is resolved. As shown in the results which follow in Section 2.5, use of the maximum likelihood estimates yields triefficiencies different to those previously published, and demonstrates that the one-wild is in fact the most critical of the three corners.

2.1.3 Scale and its estimation

Scale is a somewhat vague concept, perhaps primarily because its definition depends on the distribution in mind.

one-wild	100% slash efficiency			90% slash efficiency			80% slash efficiency		
efficiency	normal	one-wild	slash	normal	one-wild	slash	normal	one-wild	slash
100%	13	1	3	11	0	6	7	0	10
95	12	3	2	11	0	6	7	0	10
90	7	8	2	7	6	4	6	1	10
85	5	10	2	5	10	2	5	3	9
80	0	15	2	0	15	2	0	13	5
75	0	15	2	0	15	2	0	15	2

Table 2.2. Effect of understated efficiencies on triefficiency on the results given in Lax (1895). The figures given in the table are the number of estimators out of 17 that have minimum efficiency at the stated corner distribution for the specified relative efficiencies of Lax's minimum variance estimator to the true minimum variance estimator at the one-wild and slash corners.

Definition 2.5 (Location-scale family) *The "location-scale" family of distributions have a location parameter θ and a scale parameter σ (not necessarily standard deviation), and have density functions which can be written*

$$f(x) = \frac{1}{\sigma} f_0\left(\frac{x-\theta}{\sigma}\right)$$

where $f_0(x)$ depends neither on θ nor σ, and is itself a proper probability density function.

This provides a definition of scale for random variables belonging to this family. We immediately notice that if σ is a scale parameter, then for any $k > 0$, $k\sigma$ is also a scale parameter for the same family. Consider the case of the normal distribution, we have

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) = \frac{1}{\sigma'} f_0\left(\frac{x-\mu}{\sigma'}\right)$$

where $f_0(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$, and $\sigma' = \sqrt{2}\sigma$ is the scale parameter. Indeed, $f_0(x)$ is the density function of an $\mathcal{N}(0, \frac{1}{2})$ random variable. Furthermore, this choice of scale parameter satisfies all the conditions required, however it is contrary to our usual definition of the standard deviation σ as the scale parameter for the normal distribution, an assumption which facilitates $f_0(x) = \phi(x)$, the standard normal density function.

Many common continuous distributions belong to the location-scale family. A non-exhaustive list of these includes the Cauchy, slash, Student's t, exponential, double exponential (Laplace) and logistic distributions. The contaminated normal distribution CN$(p; k)$, with density function

$$f_X(x) = (1-p)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) + p\frac{1}{k\sigma}\phi\left(\frac{x-\mu}{k\sigma}\right) \tag{2.3}$$

is also a member of the location-scale family, however this is due to the particular parameterisation of the mixture distribution used here. In this case we have

$$f_0(x) = (1-p)\phi(x) + \frac{p}{k}\phi\left(\frac{x}{k}\right).$$

However for a more general mixture of two normals, in particular with $\mu_0 \neq \mu_1$, $f_0(x)$ will depend on $\mu_1 - \mu_0$ and so the distribution will not be a member of the location-scale family. Since the one-wild "distribution" does not really exist, we cannot include it in the location-scale family, however, we note that any observation in this sample is either $\mathcal{N}(\mu, \sigma^2)$ if not wild, or $\mathcal{N}(\mu, 100\sigma^2)$ if wild, and that both these distributions do qualify.

Such a general definition of scale leads to a similarly general definition of a scale estimator.

Definition 2.6 (Scale estimator) *A scale estimator for the random vector* $\mathbf{X} = (X_1, \ldots, X_n)$ *and constants* a *and* b*, is any function* $S(\mathbf{X})$ *which satisfies*

$$S(a + b\mathbf{X}) = |b|\, S(\mathbf{X}) \geq 0 \tag{2.4}$$

with equality only when all the elements of $\mathbf{X}$ *are equal.*

The most commonly used scale estimator is the sample standard deviation.

Definition 2.7 (Sample standard deviation) *The sample standard deviation for observations* $\mathbf{X} = (X_1, \ldots, X_n)$ *is given by*

$$s(\mathbf{X}) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \tag{2.5}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *is the sample mean.*

The sample standard deviation is closely related to the sample variance $s^2(\mathbf{X})$, and the latter is the minimum variance unbiased estimator for the variance parameter σ^2 for Gaussian data. Despite these excellent qualities for well-behaved data, and the fact that the sample variance will be unbiased for the underlying variance generally for random samples (where this variance is defined), experience tells us that this estimator is not robust: it is affected greatly by single outliers. Iglewicz (2000) provides an example of data of which $n-1$ observations are equal to y, and the

remaining observation is $y + na$, where $a > 0$. The sample standard deviation of this data is $a\sqrt{n}$, this depends primarily on the size of the outlying value, and is unbounded as $n \to \infty$.

The most commonly calculated robust scale estimator is probably the interquartile range (IQR) which measures the difference between a distribution's upper and lower quartiles.

Definition 2.8 (Interquartile range) *For a continuous random variable X, with cumulative distribution function $F_X(x)$, the interquartile range (IQR) is*

$$IQR(X) = F_X^{-1}(0.75) - F_X^{-1}(0.25) \tag{2.6}$$

where $x = F_X^{-1}(y)$ solves the equation $y = F_X(x)$.

For a collection of observations, we can think of F as being an empirical cumulative distribution function (cdf), and use the same rule. However, because the empirical cdf has jumps, there are conflicting methods of finding the IQR for sample data, and these are discussed in Section 2.3. The conflict arises from an attempt to make the choice of $F^{-1}(0.75)$ and $F^{-1}(0.25)$ as simple as possible. Calculating the IQR for Iglewicz's data described above, and assuming $n > 4$, we find $\text{IQR}(y, \ldots, y, y+na) = 0$ which displays the IQR's resistance to the outlying value.

Aside from the IQR, robust scale estimators are not widely used in statistical applications. The remainder of this section refers to estimators that are not formally defined until Section 2.3, but their inclusion in the discussion here serves to highlight the obscurity of these techniques. A survey of commonly used statistical software identifies very few robust scale estimators. The exceptions to this are S-PLUS (see for example Venables & Ripley 1999) and `R` (Ihaka & Gentleman 1996) which not only have many robust estimation procedures in their libraries, but also provide a simple framework for programming of additional estimators. Microsoft Excel has a built in function to compute a trimmed mean, and one to find sample percentiles, but does not include any robust scale estimation procedure. SPSS (Version 10.0.5) reports prespecified robust estimates of location for data (M-estimates based on various weight functions). However, apart from the interquartile range, SPSS does not appear to provide robust scale estimates. SHAZAM (Version 9) includes the IQR in a report of descriptive statistics; however it is not part of the default report. The Statistics Toolbox for MATLAB (Version 6.1) includes the IQR and either the

mean absolute deviation from the mean, or the MAD. There is some conflict in the on-line documentation over which estimator is calculated, although evidence seems to suggest the former non-robust estimator. SAS (Version 7) includes a relatively healthy list of robust estimators of scale: the interquartile range, Gini's mean difference, the median absolute deviation, and Rousseeuw & Croux's (1993) S_n and Q_n. While all of the above software allow users to program their own robust estimators, with the exception of S-PLUS, `R` and to a lesser extent SAS, no overt effort is made to accommodate or promote robust scale estimation.

The most comprehensive analysis of robust estimation of scale appears to be the study of Lax (1985), which follows in the wake of the Princeton Robustness Study (Andrews, Bickel, Hampel, Huber, Rogers & Tukey 1972) of location estimators. Lax considers a number of scale estimators evaluated for samples of twenty observations from Tukey's three corners. Lax considers a number of different estimators, based either on their performance in a pilot study of over 150 estimators, or on their popularity at the time. His table of efficiencies and triefficiencies is reproduced in Table 2.3. A number of estimators shown therein are dominated, i.e., at least one other estimator outperforms that estimator in all three distributions. The best of the estimators considered is the A-estimator with the biweight ψ-function and scaling constant $c = 9$. This estimator has sampling variance 85.8% of that of the best performing estimator for the one-wild distribution, and does marginally better for the normal and slash distributions. Not only does this particular class of estimators perform well in Lax's study, but it is the scale counterpart of the best performing estimator in similar studies on robust estimation of location. As we let the scaling constant $c \to \infty$, the A-estimator converges to the sample standard deviation. However, as this happens, its robustness properties are lost and its performance on long-tailed data, e.g. from the slash distribution, diminishes. Hence the choice of scaling constant $c = 9$ reflects a compromise between high efficiency in these two extreme cases.

2.2 Estimation of location and scale using the EM algorithm

The EM algorithm (Dempster et al. 1977) is a numerical method which can be used to obtain maximum likelihood (ML) estimates under situations where usual

		Efficiency		
Estimator	Normal	One-Wild	Slash	Triefficiency
A-Estimators (ψ-function)				
Biweight ($c = 6$)	65.2	77.1	90.1	65.2
Biweight ($c = 7$)	74.8	82.9	89.3	74.8
Biweight ($c = 8$)	81.8	85.4	87.6	81.8
Biweight ($c = 9$)	86.7	85.8	86.1	85.8
Biweight ($c = 10$)	90.0	84.8	84.6	84.6
Modified biweight ($c = 6$)	47.5	56.8	96.8	47.5
Sine ($c = 2.1$)	77.5	83.7	88.4	77.5
Modified sine ($c = 2.1$)	82.1	89.6	94.5	82.1
M-Estimators (Huber ψ-function)				
$b = 1.4$ (iterated)	48.1	56.8	100.0	48.1
$b = 1.7$ (iterated)	72.3	83.8	83.8	72.3
$b = 1.4$ (one-step)	55.2	68.1	86.8	55.2
$b = 1.7$ (one-step)	60.5	71.8	83.1	60.5
$b = 2.0$ (one-step)	69.8	76.1	75.9	69.8
Sample standard deviation	100.0	10.9	-	-
Trimmed standard deviation	89.9	100.0	28.1	28.1
MAD	35.3	41.5	91.8	35.3
Gaussian skip	54.7	59.3	90.1	54.7

Table 2.3. Efficiencies for selected estimators reported in Lax (1985), using Monte Carlo estimates in sample sizes of twenty. Modified biweight, sine, and modified sine A-estimators, M-estimators with the Huber ψ-function, and the Gaussian skip are defined in Lax (1985); other estimators are defined in Section 2.3. Efficiency is calculated using the sample variance of the log estimates, and is the ratio of the best performing estimator's variance to the variance of the estimator of interest. Triefficiency is the subject of Definition 2.4.

ML optimisation techniques can be problematic. There is much literature devoted to extensions and applications of the EM algorithm, and this is summarised in McLachlan & Krishnan (1997). We present a brief introduction to the EM algorithm, provide an example of its implementation for the mixture distribution $\mathrm{CN}(\frac{1}{n}; k)$, and then derive specific results for application to maximum likelihood estimation of (location and) scale for one-wild and slash samples.

We assume the observed data $\mathbf{X} = (X_1, \ldots, X_n)$ may depend on some unobserved data $\mathbf{S} = (S_1, \ldots, S_n)$. If $\mathbf{X}$ has the joint probability density function (pdf) $f(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a vector of unknown parameters, then the maximum likelihood estimator of $\boldsymbol{\theta}$ maximises the likelihood function

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}; \mathbf{X}) = f(\mathbf{x}; \boldsymbol{\theta}).$$

Defining $\mathbf{Y} = (\mathbf{X}, \mathbf{S})$ to be the complete data, we have (for discrete $\mathbf{S}$)

$$L(\boldsymbol{\theta}) = \sum_{\mathbf{S}} L_c(\boldsymbol{\theta})$$

where $L_c(\boldsymbol{\theta}) \equiv L_c(\boldsymbol{\theta}; \mathbf{X}, \mathbf{S}) = f_{\mathbf{X},\mathbf{S}}(\mathbf{x}, \mathbf{s})$ is the complete likelihood. It is often the case that the incomplete likelihood function is difficult to maximise. However, a suitable choice of $\mathbf{S}$ can facilitate a much simpler problem, and it is this property that governs the choice of $\mathbf{S}$.

The EM algorithm consists of two steps at each iteration: the first being the *Expectation* step, in which the expectation

$$Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_0\right) = \mathrm{E}_0 \left\{ \ln L_c(\boldsymbol{\theta}) | \mathbf{X} \right\}$$

is evaluated, where E_0 denotes expectation conditional on the previous estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_0$. The second step is the *Maximisation* step, in which we choose $\hat{\boldsymbol{\theta}}_1$ to maximise $Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_0\right)$ over $\boldsymbol{\theta}$. This process is then iterated to convergence. At the $(k+1)$th iteration, we calculate

$$Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k\right) = \mathrm{E}_k \left\{ \ln L_c(\boldsymbol{\theta}) | \mathbf{X} \right\}$$

(the E-step), and then choose $\hat{\boldsymbol{\theta}}_{k+1}$ such that

$$Q\left(\hat{\boldsymbol{\theta}}_{k+1}; \hat{\boldsymbol{\theta}}_k\right) > Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k\right)$$

for any other possible $\boldsymbol{\theta}$ (the M-step).

The philosophy behind and benefits of the EM algorithm are discussed by McLachlan & Krishnan (1997), however it is useful to note here some of the advantages of this method.

- The true (incomplete) likelihood cannot decrease with an additional iteration, i.e. $L(\hat{\boldsymbol{\theta}}_{k+1}) \geq L(\hat{\boldsymbol{\theta}}_k)$.
- The EM algorithm has reliable global convergence properties.
- In all of the examples considered in this thesis, at each iteration, the M-step requires no numerical maximisation and is a closed form function of the observed data $\mathbf{X}$ and the previous parameter estimate $\hat{\boldsymbol{\theta}}_k$.
- Although the EM algorithm can be slow to converge, for the examples considered in this thesis, each iteration of the algorithm proceeds with low computing cost offsetting the slow convergence effect.

Use of the EM algorithm is demonstrated in the following section.

2.2.1 General results, and an example

In this section, we consider observations $\mathbf{X} = (X_1, \ldots, X_n)$ where $\mathbf{X}$ depends on unobserved data $\mathbf{S} = (S_1, \ldots, S_n)$. The particular construction we adopt will not only be useful for the three corner distributions considered in the simulation study that follows, but also for the Student's t distributions, and the contaminated normal $\mathrm{CN}(p; k)$.

Definition 2.9 (Gaussian compound scale model) *Observations $X_1, \ldots, X_n$ are said to follow a Gaussian compound scale model with parameters μ and σ^2 if, given $\mathbf{S} = (S_1, \ldots, S_n)$, the X_i are independent $\mathcal{N}(\mu, \sigma^2/S_i)$ random variables, where the S_i are non-negative with known distribution.*

It follows from the above definition that we can write

$$X_i = \mu + \sigma \frac{Z_i}{\sqrt{S_i}} \qquad (i = 1, \ldots, n) \tag{2.7}$$

where the Z_i are independent standard normal random variables. As a consequence, the compound normal sample is sometimes referred to as the normal/independent sample, describing the ratio of a normal variable to a general, independently distributed random variable.

A random sample from the slash distribution clearly falls in this category. The numerator of the slash is indeed a standard normal, and the denominator is a uniform

random variable independent of the normal. The Student's t-distribution is also of this sort, where the denominator is the square root of a Chi-squared random variable divided by its degrees of freedom. In the one-wild case, the X_i are not independent, since there is only a single "wild" observation and knowledge of which observation this is, has an impact on the other observations. Nonetheless, we note that, conditional on the S_i, the X_i *are* independent due to the independence of the Z_i, and the one-wild sample follows a Gaussian compound scale model.

We wish to find the maximum likelihood estimates of μ and σ^2 based only on the observations $\mathbf{X} = (X_1, \ldots, X_n)$ and now use the EM algorithm (Dempster et al. 1977) to construct an iterative formula for estimating these parameters. We note that $\mathbf{X}$ is the incomplete data, and $(\mathbf{X}, \mathbf{S})$ is the complete data. Due to the choice of $\mathbf{S}$ in Definition 2.9, we will show that finding the ML estimates of μ and σ^2 is relatively straightforward for data from a Gaussian compound scale model.

Theorem 2.1 *The maximum likelihood estimators of μ and σ^2 for observations $\mathbf{X} = (X_1, \ldots, X_n)$ following a Gaussian compound scale model, are found by iterating the equations*

$$\hat{\mu} = \frac{\sum_{i=1}^{n} E_0(S_i|\mathbf{X}) X_i}{\sum_{i=1}^{n} E_0(S_i|\mathbf{X})} \tag{2.8}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} E_0(S_i|\mathbf{X})(X_i - \hat{\mu})^2 \tag{2.9}$$

where $E_0(S_i|\mathbf{X})$ is the expectation of S_i given $\mathbf{X}$, evaluated at previous estimates of μ and σ^2.

Proof Conditional on S_i, it follows from Definition 2.9 that the X_i are independent normal random variables with mean μ and variance σ^2/S_i, and thus the joint distribution of $\mathbf{X}$ and $\mathbf{S} = (S_1, \ldots, S_n)$ is given by

$$dF_{\mathbf{X},\mathbf{S}}(\mathbf{x}, \mathbf{s}) = f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s}) d\mathbf{x} dF_{\mathbf{S}}(\mathbf{s}) = \left[\prod_{i=1}^{n} \frac{\sqrt{s_i}}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}u_i^2 s_i} \right] dF_{\mathbf{S}}(\mathbf{s}) d\mathbf{x}$$

where $U_i = (X_i - \mu)/\sigma$ is the standardised score, u_i its realisation and $F_{\mathbf{S}}(\mathbf{s})$ is the distribution function of $\mathbf{S}$. Consequently the complete log-likelihood is given by

$$\ln L_c(\mu, \sigma^2) = \sum_{i=1}^{n} \left(-\frac{1}{2} \ln \sigma^2 - \frac{1}{2} U_i^2 S_i \right) + \text{constant}$$

where all terms not featuring μ or σ^2 are included in the constant. The theory behind the EM algorithm leads us to maximise the smoothed likelihood

$$Q(\mu, \sigma^2; \hat{\mu}_0, \hat{\sigma}_0^2) = \mathrm{E}_0 \left[\frac{2}{n} \sum_{i=1}^{n} \left(-\frac{1}{2} \ln \sigma^2 - \frac{1}{2} \frac{S_i (X_i - \mu)^2}{\sigma^2} \right) \middle| \mathbf{X} \right]$$

$$= -\ln \sigma^2 - \frac{1}{n\sigma^2} \sum_{i=1}^{n} \mathrm{E}_0(S_i|\mathbf{X})(X_i - \mu)^2 \qquad (2.10)$$

with respect to both μ and σ^2, where E_0 denotes expectation over the conditional distribution using the estimates $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ in place of the true μ and σ^2. Maximising (2.10) with respect to μ requires solution of the equation

$$-\frac{1}{n\sigma^2} \sum_{i=1}^{n} \mathrm{E}_0(S_i|\mathbf{X})(-2)(X_i - \mu) \Big|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} = 0$$

which yields

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \mathrm{E}_0(S_i|\mathbf{X}) X_i}{\sum_{i=1}^{n} \mathrm{E}_0(S_i|\mathbf{X})}$$

which is a function of $\mathbf{X}$ and $\hat{\boldsymbol{\theta}}_0 = (\hat{\mu}_0, \hat{\sigma}_0^2)$ alone. Maximising (2.10) with respect to σ^2 requires solution of the equation

$$-\frac{1}{\sigma^2} + \frac{1}{n\sigma^4} \sum_{i=1}^{n} \mathrm{E}_0(S_i|\mathbf{X})(X_i - \mu)^2 \Big|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} = 0$$

which yields

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_0(S_i|\mathbf{X})(X_i - \hat{\mu})^2$$

as required. □

In the degenerate case where the X_i are normal, and $S_i = 1$ for all i, clearly $\mathrm{E}_0(S_i|\mathbf{X}) = 1$ for all i, and we obtain $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$ the sample mean, and $\hat{\sigma}^2 = \frac{n-1}{n} s^2(\mathbf{X})$ the familiar maximum likelihood estimator of variance.

Dempster et al. (1977) note the similarity between the iterations given in Theorem 2.1 and iteratively reweighted least squares. The form for the updated $\hat{\mu}$ and $\hat{\sigma}^2$ is of a weighted average of the observations and squared deviations respectively, with the weights dependent on the parameter estimates from the previous iteration (Dempster et al. 1977, Section 4.6).

The following theorem may assist in the evaluation of $\mathrm{E}_0(S_i|\mathbf{X})$.

Theorem 2.2 *If the observations* $\mathbf{X} = (X_1, \ldots, X_n)$ *follow the Gaussian compound scale model given in Definition 2.9, and if in addition the* S_i *are independent of one another with distribution function* $F_{S_i}(s)$*, then* $E_0(S_i|\mathbf{X})$ *in Theorem 2.1 is given by*

$$E_0(S_i|\mathbf{X}) = G\left(\frac{1}{2}\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2\right)$$

where $G(t) = -\frac{d}{dt}\ln M(t)$,

$$M(t) = \int_{s=0}^{\infty} e^{-ts} s^{\frac{1}{2}} dF_{S_i}(s)$$

and $\hat{\mu}_0$ *and* $\hat{\sigma}_0^2$ *are previous estimates of* μ *and* σ^2 *respectively.*

Proof Since the S_i are independent for $i = 1, \ldots, n$, it follows from Definition 2.9 that the X_i are independent. Hence

$$E_0(S_i|\mathbf{X}) = E_0(S_i|X_i)$$

and, dropping the subscripts,

$$dF_{S|X}(s|x) \propto f_{X|S}(x|s)dx dF_S(s) \propto \sqrt{s}\exp\left(-\frac{1}{2}\frac{s(x-\mu)^2}{\sigma^2}\right) dF_S(s)dx$$

since given S_i, X_i is normal with mean μ and variance σ^2/S_i. Consequently

$$\mathrm{E}_0(S_i|\mathbf{X}) = \frac{\int_{s=0}^{\infty} e^{-\frac{1}{2}sU_i^2} s^{\frac{3}{2}} dF_{S_i}(s)}{\int_{s=0}^{\infty} e^{-\frac{1}{2}sU_i^2} s^{\frac{1}{2}} dF_{S_i}(s)}$$

where $U_i = (X_i - \hat{\mu}_0)/\hat{\sigma}_0$ and where the denominator is the normalising constant. Define the Laplace transform

$$M(t) = \int_{s=0}^{\infty} e^{-ts} s^{\frac{1}{2}} dF_{S_i}(s) \qquad (t \geq 0)$$

so that

$$M'(t) = -\int_{s=0}^{\infty} e^{-ts} s^{\frac{3}{2}} dF_{S_i}(s)$$

and now

$$\mathrm{E}_0(S_i|\mathbf{X}) = \frac{-M'\left(\frac{1}{2}U_i^2\right)}{M\left(\frac{1}{2}U_i^2\right)} = G(t)\Big|_{t=\frac{1}{2}\left(\frac{X_i-\hat{\mu}_0}{\hat{\sigma}_0}\right)^2}$$

where $G(t) = -\frac{d}{dt}\ln M(t)$. □

In order to give better insight into the implementation and performance of the EM algorithm, we demonstrate the application of the EM algorithm to a random

sample from the contaminated normal distribution $\mathrm{CN}(\frac{1}{n}; k)$ for $n = 10$ and $k = 10$. A sample from this distribution may be one-wild, however, as discussed earlier, the actual number of observations drawn from the "wild" distribution has a Binomial distribution.

To apply the EM algorithm to a sample from the mixture distribution, we must first evaluate $\mathrm{E}_0(S_i|\mathbf{X})$ in this case. The necessary result is given in the following theorem.

Theorem 2.3 *The maximum likelihood estimators of location and scale for a random sample from the contaminated normal distribution $CN(\frac{1}{n}; k)$ are found by iterating equations (2.8) and (2.9) with*

$$E_0(S_i|\mathbf{X}) = 1 - \frac{(1 - \frac{1}{k^2}) \exp\left[\frac{1}{2}(1 - \frac{1}{k^2})\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2\right]}{k(n-1) + \exp\left[\frac{1}{2}(1 - \frac{1}{k^2})\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2\right]}$$

where $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ are the previous estimates of μ and σ^2 respectively, and where k is assumed known.

Proof For a random variable distributed as $\mathrm{CN}(\frac{1}{n}; k)$, the $\sqrt{S_i}$ are independent random variables that take on the value 1 with probability $1 - \frac{1}{n}$ and $\frac{1}{k^2}$ with probability $\frac{1}{n}$. Since S_i is a discrete random variable, we define

$$dF_{S_i}(s) = \begin{cases} 1 - \frac{1}{n} & s = 1 \\ \frac{1}{n} & s = \frac{1}{k^2} \\ 0 & \text{otherwise} \end{cases}$$

for all i. Since the S_i are independent, Theorem 2.2 applies, with

$$M(t) = \int_{s=0}^{\infty} e^{-ts} s^{\frac{1}{2}} dF_{S_i}(s) = (1 - \tfrac{1}{n})e^{-t} + \tfrac{1}{nk} e^{-t/k^2}.$$

Differentiating with respect to t, we find

$$G(t) = \frac{-M'(t)}{M(t)} = \frac{(1 - \frac{1}{n})e^{-t} + \frac{1}{nk^3}e^{-t/k^2}}{(1 - \frac{1}{n})e^{-t} + \frac{1}{nk}e^{-t/k^2}} = 1 - \frac{\frac{1}{nk}(1 - \frac{1}{k^2})e^{-t/k^2}}{(1 - \frac{1}{n})e^{-t} + \frac{1}{nk}e^{-t/k^2}}.$$

Multiplying through by nke^t, we obtain

$$G(t) = 1 - \frac{(1 - \frac{1}{k^2})e^{(1-\frac{1}{k^2})t}}{k(n-1) + e^{(1-\frac{1}{k^2})t}}$$

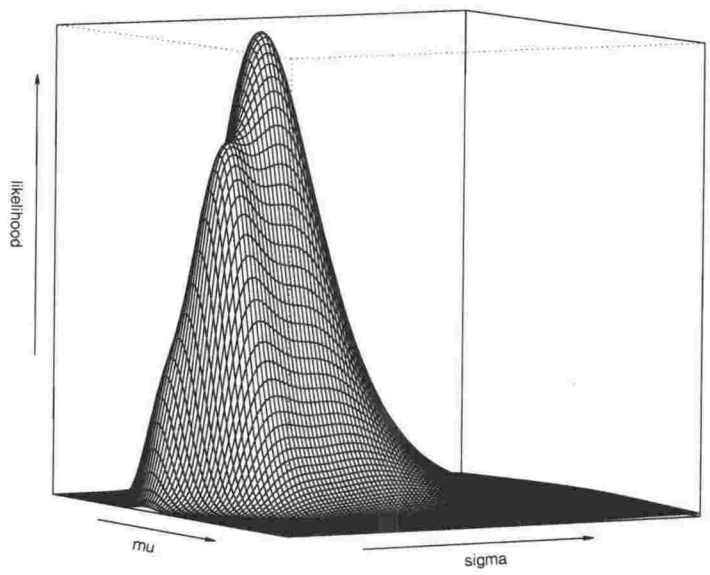

Figure 2.4. The likelihood function for the sample of size ten from the CN($\frac{1}{10}$, 10) distribution described in Table 2.4. The likelihood function is given in (2.11), and is plotted against a grid in which $-1 \leq \mu \leq 1$ and $0.2 \leq \sigma \leq 2$.

and we evaluate this function at $t = \frac{1}{2}\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2$ as required. □

The density function for the contaminated normal random variable was given in (2.3) and thus the (incomplete) likelihood for the sample is given by

$$L(\boldsymbol{\theta}) = f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{n-1}{n\sigma} \phi\left(\frac{x_i - \mu}{\sigma}\right) + \frac{1}{nk\sigma} \phi\left(\frac{x_i - \mu}{k\sigma}\right) \right] \tag{2.11}$$

where $\boldsymbol{\theta} = (\mu, \sigma^2)$ are the unknown parameters, and k is assumed known. In the case where $n = 10$ and $k = 10$, the maximum likelihood estimates of μ and σ^2 are given by Theorem 2.1, with

$$\mathrm{E}_0(S_i|\mathbf{X}) = 1 - \frac{0.99 \exp\left[0.495\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2\right]}{90 + \exp\left[0.495\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2\right]}$$

where $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ are the previous estimates of μ and σ^2 respectively.

The EM algorithm is applied to ten randomly sampled observations from the contaminated normal distribution CN($\frac{1}{10}$, 10) with $\mu = 0$ and $\sigma = 1$. The sample itself is given in Table 2.4 and the likelihood function of this sample is shown in Figure 2.4 for $-1 \leq \mu \leq 1$ and $0.2 \leq \sigma \leq 2$. This likelihood function has a maximum at $\mu = -0.341$ and $\sigma = 0.756$.

Table 2.4 describes the evolution of the EM estimates for this sample from the mixture. The sample is ordered in the table, which enables easier comparison of

		Iteration					
i	$x_{(i)}$	1	2	3	4	Final	One-wild
1	-2.184	0.978	0.975	0.969	0.910	0.827	1.000
2	-0.975	0.987	0.987	0.987	0.986	0.985	1.000
3	-0.561	0.988	0.988	0.989	0.989	0.989	1.000
4	-0.445	0.989	0.989	0.989	0.989	0.989	1.000
5	-0.355	0.989	0.989	0.989	0.989	0.989	1.000
6	-0.023	0.989	0.989	0.989	0.988	0.988	1.000
7	0.090	0.989	0.989	0.989	0.988	0.987	1.000
8	0.100	0.989	0.989	0.989	0.988	0.987	1.000
9	0.970	0.988	0.988	0.986	0.968	0.954	1.000
10	5.429	0.763	0.437	0.053	0.010	0.010	0.010
$\hat{\mu}$	0.205	0.085	-0.101	-0.338	-0.356	-0.341	-0.370
$\hat{\sigma}$	2.010	1.721	1.413	0.877	0.776	0.756	0.799

Table 2.4. Maximum likelihood estimation of μ and σ for a simulated sample of size ten from the $\mathrm{CN}(\frac{1}{10}, 10)$ distribution with $\mu = 0$ and $\sigma = 1$. The main entries of the table are the values of $\mathrm{E}_0(S_i|\mathbf{X})$ for the ordered sample at the indicated iteration. The $x_{(i)}$ are the ordered sample observations, and the final column gives the corresponding values of $\mathrm{E}_0(S_i|\mathbf{X})$ if the sample is assumed one-wild. The final two rows of the table give the estimates of $\hat{\mu}$ and $\hat{\sigma}$ at the indicated iteration and for the one-wild fit. The initial values of $\hat{\mu}$ and $\hat{\sigma}$, given below the data, are the sample mean and standard deviation respectively.

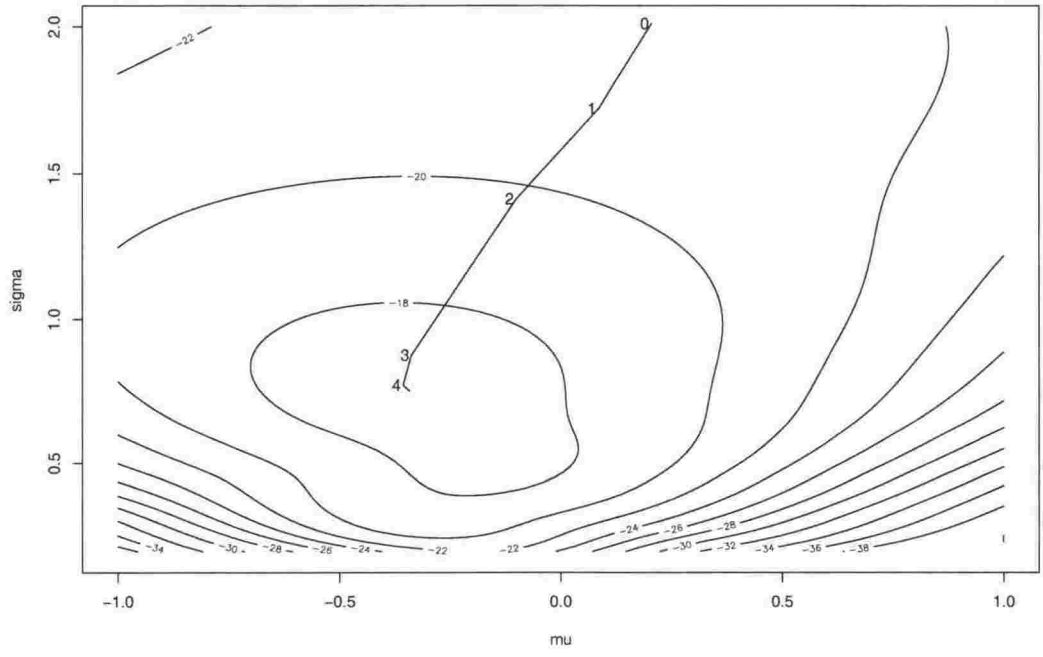

Figure 2.5. A contour plot of the log-likelihood function corresponding to the likelihood function shown in Figure 2.4. The log-likelihood is constant along each contour. In addition, the path of $(\hat{\mu}, \hat{\sigma})$ is shown and labeled by the iteration number of the EM algorithm.

the "weights" $\mathrm{E}_0(S_i|\mathbf{X})$ given to each observation at each iteration. By chance, the sample also happens to be one-wild: there are nine observations drawn from the standard normal distribution, and the tenth is from the contaminating distribution $\mathcal{N}(0, 100)$. The EM iterations are initialised using the sample mean $\bar{x}$ and standard deviation s, and these are given in the table directly below the observations. At the first iteration, the two most extreme values are identified and given lower weights than the other observations, however as the recursion proceeds and $\hat{\sigma}$ decreases, the outlier (5.429) is clearly identified as a contaminated observation, and given minimum weight $\frac{1}{100}$. The other extreme observation is treated with caution, and gets a weight smaller than the remaining observations. From the fourth iteration to the final iteration there is very little change in the weights or the parameter estimates.

The final weights for the one-wild sample (whose form is given in Theorem 2.5 which follows) are also given in Table 2.4 for comparison. Unlike for a sample from the contaminated normal, in this case, it is known that there is only a single "wild" observation. Thus, the most extreme observation is identified, and because of its size, it is the only candidate for an outlier. It gets the minimum weight, and the other observations full weight, and results in treatment for the data that is identical to what might be performed manually: identify the outlier and downweight it, and use the sample mean and standard deviation. Unlike the mixture recursions, the one-wild recursions need only four iterations for convergence, as they have the additional information that only a single value is "wild".

The progress of the estimates is shown graphically in Figure 2.5. Fixed contours of the log-likelihood are shown, as well as the trace of $\hat{\boldsymbol{\theta}}_k$ for $k = 0, \ldots, 15$. The initial values $(\bar{x}, s)$ are labeled '0', and the updated estimates from the first iteration of the EM algorithm labeled '1', etc. After the fourth iteration, convergence has almost been achieved, and so the remaining eleven estimates are not labeled. This plot confirms the convergence of the estimates to the maximum likelihood estimates, and also the monotonicity of the $L(\hat{\boldsymbol{\theta}}_k)$ sequence.

2.2.2 Maximum likelihood scale estimation for the one-wild and slash

In Lax's study, the minimum variance estimators of scale for the one-wild and slash distributions were both weighted averages which gave zero weight to the most extreme observations. However, no theory was provided to support this result, nor

does it seem plausible, particularly for the one-wild. In order to secure the results of this study, it is important to have the minimum variance estimators in all three situations: normal, one-wild and slash. We use the asymptotically efficient maximum likelihood estimator in each case and hope that the finite sample properties of these estimators allow them to be close-to-optimal in the one-wild and slash cases. In particular, we optimise the likelihood of the sample by choice of both location and scale estimates. The former is included to be consistent with use of the sample mean as an auxiliary location estimator for the normal situation, and so the scale estimates are neither based on the knowledge that the data have theoretical location zero, nor based on a sub-optimal location estimate like the sample median. Although the maximum likelihood estimators are asymptotically efficient but not necessarily efficient for finite samples, there is the possibility that some estimators may have efficiency greater than 100%.

For normal data, with $S_i = 1$ for each i, the maximum likelihood location estimate is the sample mean and the maximum likelihood scale estimate is proportional to the sample standard deviation, with no iteration required.

The slash is the ratio of an $\mathcal{N}(\mu, \sigma^2)$ random variable and a uniform random variable on the interval $[0, 1]$, and has density function

$$f_X(x) = \begin{cases} \frac{\sigma}{(x-\mu)^2\sqrt{2\pi}} \left[1 - \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)\right] & x \neq 0 \\ \frac{1}{2\sigma\sqrt{2\pi}} & x = 0. \end{cases}$$

The maximum likelihood estimators for the slash parameters are known (see Kafadar 1982) but are confirmed here using the EM technology introduced in Section 2.2.1.

Theorem 2.4 *The maximum likelihood estimators of location and scale for a random sample* $\mathbf{X} = (X_1, \ldots, X_n)$ *from the slash distribution are found by iterating equations (2.8) and (2.9) with*

$$E_0(S_i|\mathbf{X}) = \frac{2\hat{\sigma}_0^2}{(X_i - \hat{\mu}_0)^2} - \left[\exp\left(\frac{1}{2}\frac{(X_i - \hat{\mu}_0)^2}{\hat{\sigma}_0^2}\right) - 1\right]^{-1} \tag{2.12}$$

where $\hat{\mu}_0$ *and* $\hat{\sigma}_0^2$ *are the previous estimates of* μ *and* σ^2 *respectively.*

Proof For the slash distribution, the $\sqrt{S_i}$ are independent uniform random variables on the interval $[0, 1]$, and hence the conditions of Theorem 2.1 are met. The distribution function of S_i is

$$F_{S_i}(s) = \Pr(S_i < s) = \Pr(U_i < \sqrt{s}) = \begin{cases} 1 & s > 0 \\ \sqrt{s} & 0 < s < 1 \\ 0 & s < 0 \end{cases}$$

where U_i is uniform on the interval $[0, 1]$ and the probability density function of S_i is

$$f_{S_i}(s) = \begin{cases} \frac{1}{2}s^{-\frac{1}{2}} & 0 < s < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Since the S_i are independent, Theorem 2.2 applies, with

$$M(t) = \int_{s=0}^{\infty} e^{-ts} s^{\frac{1}{2}} dF_{S_i}(s) = \int_{s=0}^{1} e^{-ts} ds = t(1 - e^{-t}).$$

Differentiating with respect to t, we find

$$G(t) = \frac{-M'(t)}{M(t)} = \frac{(1 - e^{-t}) + te^{-t}}{t(1 - e^{-t})} = \frac{1}{t} + \frac{1}{e^t - 1}$$

and we evaluate this function at $t = \frac{1}{2}\left(\frac{(X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2$ as required. □

Application of Theorems 2.1 and 2.4 for data from the slash distribution yields the maximum likelihood estimators of μ and σ^2, which we will use to provide the minimum variance estimates, and to form the basis of comparison for the performance of scale estimators for slash data. In this particular case the equations following from Theorem 2.4 are identical to those given by Kafadar (1982) using traditional maximum likelihood techniques. Once converged, through maximum likelihood theory, this method provides the asymptotic minimum variance estimator of scale for the slash distribution. This estimator is not proposed as one which is likely to be useful in general, and hence we compute it only for the slash distribution. The EM recursions are favoured over traditional maximum likelihood techniques due to their desirable computational properties.

The above analysis is also possible for the one-wild distribution, although in this case the S_i are not independent and thus Theorem 2.2 does not apply. Kafadar (1982) states that the one-wild sample is not a sample from any particular distribution, and that no maximum likelihood method is helpful. However the parameters of the one-wild do indeed have maximum likelihood estimates, and the EM algorithm yields these. The relevant weight function $\mathrm{E}_0(S_i|\mathbf{X})$ is given in the following theorem.

Theorem 2.5 *The maximum likelihood estimators of location and scale for a one-wild sample are found by iterating equations (2.8) and (2.9) with*

$$E_0(S_i|\mathbf{X}) = 1 - \frac{99}{100} \frac{\exp\left[0.495\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2\right]}{\sum_{i=1}^{n} \exp\left[0.495\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2\right]}$$

where $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ are the previous estimates of μ and σ^2 respectively.

Proof Here Theorem 2.1 applies with

$$S_i = \begin{cases} \frac{1}{100} & i = N \\ 1 & i \neq N. \end{cases}$$

where N is a discrete uniform random variable with

$$P(N = i) = \frac{1}{n} \qquad (i = 1, \ldots, n).$$

For any sample, $1 \leq N \leq n$ is drawn, and this observation is the "wild" observation. For $i \neq N$, $S_i = 1$ and so $X_i \sim \mathcal{N}(\mu, \sigma^2)$, whereas for $i = N$, $S_i = \frac{1}{100}$ and so $X_i \sim \mathcal{N}(\mu, 100\sigma^2)$ as required.

Now consider the joint stochastic properties of $\mathbf{X}$ and $\mathbf{S}$, which is equivalent to considering the joint stochastic properties of $\mathbf{X}$ and N. Conditioning, we see

$$\begin{aligned} dF_{\mathbf{X},\mathbf{S}}(\mathbf{x}, \mathbf{s}) &= dF_{\mathbf{X},N}(\mathbf{x}, i) = f_{\mathbf{X}|N}(\mathbf{x}|i)P(N = i)d\mathbf{x} \\ &= \frac{1}{n}\left[\prod_{j=1, j\neq i}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(X_j-\mu)^2}{\sigma^2}}\right] \frac{1}{10\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(X_i-\mu)^2}{100\sigma^2}} d\mathbf{x} \\ &\propto \exp\left(\frac{1}{2}\frac{99}{100}U_i^2\right) \end{aligned}$$

where $U_i = (X_i - \mu)/\sigma$ is the standardised score. Thus

$$P(N = i|\mathbf{X}) = \frac{\exp\left(\frac{1}{2}\frac{99}{100}\frac{(X_i-\mu)^2}{\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(\frac{1}{2}\frac{99}{100}\frac{(X_i-\mu)^2}{\sigma^2}\right)} \qquad (i = 1, \ldots, n)$$

with the denominator ensuring $P(N = i|\mathbf{X})$ is a proper probability function. Thus we determine

$$\begin{aligned} \mathrm{E}_0(S_i|\mathbf{X}) &= \frac{1}{100}P(N = i|\mathbf{X}, \hat{\boldsymbol{\theta}}_0) + P(N \neq i|\mathbf{X}, \hat{\boldsymbol{\theta}}_0) \\ &= 1 - \frac{99}{100}\frac{\exp\left(\frac{1}{2}\frac{99}{100}\frac{(X_i-\hat{\mu}_0)^2}{\hat{\sigma}_0^2}\right)}{\sum_{i=1}^{n} \exp\left(\frac{1}{2}\frac{99}{100}\frac{(X_i-\hat{\mu}_0)^2}{\hat{\sigma}_0^2}\right)} \end{aligned}$$

as required. □

Application of Theorems 2.1 and 2.5 for a one-wild sample yields the maximum likelihood estimators of μ and σ^2, which we will use to provide the minimum variance estimates, and to form the basis of comparison for the performance of scale estimators for one-wild data. An example of this was provided in an earlier example (see Table 2.4) for comparison to ML estimation for the contaminated normal

distribution. Note that unlike the contaminated normal and the slash, the weight function for each observation in the one-wild sample depends on all n observations, rather than just the observation of interest, due to the dependent nature of the S_i.

Thus, the minimum variance estimators are secured for all three corner distributions, and the results provided in Section 2.5 do not depend on the other estimators considered. Consequently the results given also provide a benchmark for efficiency comparison with estimators not considered here, using measures such as Tukey's triefficiency.

2.3 Scale estimators

In this section, we describe the estimators included in the simulation study. They are divided into two classes. The first are described as single-pass scale estimators, since they do not require an auxiliary estimate of scale. The second class of scale estimators examined are multi-pass estimators. Here, the focus is on identifying a general purpose scale estimator, and hence the estimators depend only on two passes through the data. Exceptions are the maximum likelihood (ML) estimators, which are iterated until convergence for the one-wild and slash samples, and ML estimators for the Student's t-distributions. In each case, this ensures the true maximum likelihood estimates are calculated and allows comparison on this basis.

Most scale estimators rely on an auxiliary estimate of location, generally the sample median, which is used unless the definition of the scale estimator dictates otherwise. An example of this is the sample standard deviation, for which the sample mean is used. A comparable simulation to what follows is done for three prominent location estimators, and this is reported in Appendix B. While the results of that simulation are interesting in their own right, they do not conform to the focus of this thesis and are hence omitted from the main text with deeper analysis left for further research.

The simulations performed by Lax resulted in eight undominated estimators. An estimator is considered dominated if its efficiency at every distribution is less than the efficiency of another estimator for each of those distributions considered, i.e. it is worse than some other estimator in all instances. Referring to Table 2.3, we see that the Gaussian skip estimator, for example, is dominated by the A-estimators with the biweight ψ-function and $c = 6$, and the modified sine ψ-function with $c = 2.1$.

We omit the dominated estimators of Lax from this study, with the exception of the MAD due to its popularity.

The iterated M-estimator, with the Huber ψ-function and $b = 1.4$, was among Lax's undominated estimators; however it too will not be considered in this study. The main reason for this is Lax's comment "when one intends to use a scale estimator in an automatic fashion as part of a larger algorithm, the Huber scale estimator may be an unsuitable choice" (Lax 1985, page 739). Properties of the EM algorithm ensure that the ML estimates obtained are not subject to the same criticism, and these are the only iterated estimators included.

2.3.1 Single-pass scale estimators

Various estimators which depend only on a single pass of the data are defined below for the observations $\mathbf{X} = (X_1, \ldots, X_n)$. While some of these depend on an auxiliary estimate of location, which could be seen as a pass of the data, the location estimates are generally of a simple form and not computationally intensive.

The sample standard deviation was defined in Definition 2.7; however it can also be defined via the equation

$$s(\mathbf{X}) = \sqrt{\frac{1}{n(n-1)} \sum_{i<j} (X_i - X_j)^2} \tag{2.13}$$

which shows the sample variance is proportional to the average of the squared interpoint distances $X_i - X_j$. Outlying values will result in many interpoint distances being large, with $s(\mathbf{X})$ inflated as a result.

For any random sample drawn from a distribution with finite variance, the sample variance will be an unbiased estimator of this. It follows that the sample standard deviation is a biased estimator of the underlying population standard deviation, but this bias has an analytic expression when the data are Gaussian. The expected value of the sample standard deviation in the case of Gaussian data is

$$\mathrm{E}(s(\mathbf{X})) = \sigma \left[\left(\frac{2}{n-1} \right)^{\frac{1}{2}} \Gamma\left(\frac{1}{2} n \right) \right] \Big/ \Gamma\left(\frac{1}{2} n - \frac{1}{2} \right) \tag{2.14}$$

where $\Gamma(x)$ is the gamma function. In the case where $n = 20$, the sample standard deviation has expected value 0.9869σ for a normal sample.

Definition 2.10 (Gini's mean difference) *Gini's mean difference for the observations* $\mathbf{X} = (X_1, \ldots, X_n)$ *is given by*

$$G(\mathbf{X}) = \frac{2}{n(n-1)} \sum_{i<j} |X_i - X_j| . \tag{2.15}$$

Gini's mean difference is a similar estimator to the sample standard deviation, with the squared interpoint differences seen in (2.13) being replaced by absolute differences. For reasons outlined above for the sample standard deviation, this statistic is also not very robust, since the absolute difference between every pair of observations is computed. However, use of the absolute value rather than the square reduces the impact of large differences. This statistic forms the basis of robust estimation of risk in a strand of the financial literature. Typically return sample variance is used to quantify risk, however Shalit & Yitzhaki (1984) employ Gini's mean difference as an alternative measure. They are motivated mainly by the theoretical results it facilitates, rather than robustness.

Definition 2.11 (Trimmed sample standard deviation) *The trimmed sample standard deviation for the observations* $\mathbf{X} = (X_1, \ldots, X_n)$ *is given by*

$$s_{trim}(\mathbf{X}; p, r) = \sqrt{M_{1,r}\left((\mathbf{X} - M_{2,p}(\mathbf{X}))^2\right)} \tag{2.16}$$

where $M_{2,p}(\mathbf{X})$ *is a two-sided* $100p\%$ *trimmed mean, which takes the arithmetic average of a reduced data set, where the* $[pn/2]$ *smallest and the* $[pn/2]$ *largest observations are omitted, and where* $M_{1,r}(\mathbf{X})$ *is a one-sided* $100r\%$ *mean, which omits the largest* $[rn]$ *observations from the arithmetic average. Here* $[q]$ *denotes the integer part of* q.

Thus, the trimmed sample standard deviation alters the sample standard deviation in two ways in order to reduce the effects of outliers, and has two parameters p and r. Firstly, rather than using the sample mean as the auxiliary location estimator, the most extreme observations from each end of the ordered sample are omitted. Secondly, the squared deviations about this mean are formed, and the largest of these are omitted from the second average.

In the Lax study $p = r = 0.2$ with $n = 20$, and so in each calculation, four observations are ignored. This choice of p and r resulted in the highest efficiency for the one-wild distribution in Lax's study; however this seems suboptimal. In cases

where the "wild" observation is very large indeed, the final weights $\mathrm{E}_0(S_i|\mathbf{X})$ are unity for the "good" observations, and $\frac{1}{100}$ for the "wild" observation as seen for the example illustrated in Table 2.4. Thus the wild observation is not discarded, but down-weighted so that the weighted squared deviation behaves much like the others.

Definition 2.12 (Sample interquartile range) *The sample interquartile range for the observations* $\mathbf{X} = (X_1, \ldots, X_n)$ *is given by*

$$IQR(\mathbf{X}) = UQ(\mathbf{X}) - LQ(\mathbf{X}) \tag{2.17}$$

where $LQ(\mathbf{X})$ *is the sample lower quartile, given by*

$$LQ(\mathbf{X}) = \big(1 - (\ell^* - \ell)\big)X_{(\ell)} + (\ell^* - \ell)X_{(\ell+1)}$$

where $X_{(i)}$ *is the ith order statistic of the observations* $\mathbf{X}$, $\ell^* = 1 + \frac{1}{4}(n-1)$ *and* $\ell = [\ell^*]$, *and where* $UQ(\mathbf{X})$ *is the sample upper quartile, given by*

$$UQ(\mathbf{X}) = \big(1 - (u^* - u)\big)X_{(u)} + (u^* - u)X_{(u+1)}$$

where $u^* = 1 + \frac{3}{4}(n-1)$ *and* $u = [u^*]$.

The sample interquartile range (IQR), defined for a continuous random variable X in Definition 2.8 as the difference between the 25th and 75th percentiles of X, can be calculated for the observations $\mathbf{X} = (X_1, \ldots, X_n)$ using the empirical cdf; however this can be ambiguous. Definition 2.12 follows the technique used in `R` (Ihaka & Gentleman 1996), and provides one way of resolving this ambiguity. The sample IQR can then be used in a simple technique of outlier detection, and is chosen because it is resistant to outliers, since it ignores the most extreme 25% of each tail. As mentioned earlier, there are a range of methods used to calculate the sample lower and upper quartiles, most used because of their simplicity. The fourths, described in Hoaglin, Mosteller & Tukey (2000), are similar to the quartiles and are found using a simple algorithm. The lower fourth is given by the observation with position

$$k = \frac{[(n+1)/2] + 1}{2}$$

in the ordered data, denoted $X_{(k)}$, where $[x]$ is the integer part of x. If k is not a positive integer, then the lower fourth is the average of the observations $X_{(k)}$ and $X_{(k+1)}$. This will be compared to the lower quartile below.

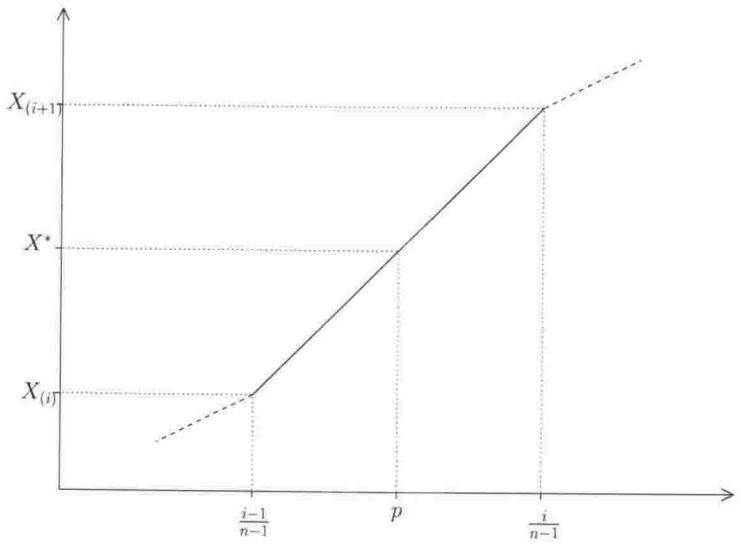

Figure 2.6. The method used by R to find a sample percentile. $X_{(i)}$ is the ith largest observation in a collection of data of size n. X^* is the pth sample percentile.

Definition 2.12 is based on the method used by R (Ihaka & Gentleman 1996) to calculate the lower and upper quartiles of a collection of data using interpolation. In particular, for any sample percentile p we define $r = 1 + (n-1)p$ and set $i = [r]$ (the integer part of r). Then the sample percentile is given by

$$\text{percentile} = \big(1 - (r-i)\big)X_{(i)} + (r-i)X_{(i+1)}.$$

This method amounts to linear interpolation of the ordered observations $X_{(1)}, \ldots, X_{(n)}$ against the sequence $0, \frac{1}{n-1}, \ldots, \frac{n-2}{n-1}, 1$. Any given percentile can be obtained from this curve as shown graphically in Figure 2.6. In the case where $n = 4q+1$ or $4q-1$ for $q \in \mathbb{Z}^+$ the lower quartile will equal the lower fourth; however, when $n = 4q$ or $4q+2$, these will not be equal in general.

Boxplots are commonly drawn with observations 1.5 times the interquartile range above the upper quartile or below the lower quartile shown as points rather than included in the whiskers (this practice is observed throughout this thesis). The points that lie outside this range are considered to be potential outliers, and for a random sample from the normal distribution with n large, we would expect only 0.7% of the observations to be labelled in this way. Further discussion of this technique can be found in Hoaglin et al. (2000).

The IQR is the simplest estimator of scale considered here, is commonly used, and is certainly the easiest to compute by hand.

Definition 2.13 (Median absolute deviation) *The median absolute deviation (MAD) for the observations* $\mathbf{X} = (X_1, \ldots, X_n)$ *is given by*

$$MAD(\mathbf{X}) = median_i \left| X_i - median_j(X_j) \right| . \tag{2.18}$$

Thus, the MAD is the median of the absolute deviations of the observations $\mathbf{X}$ about their median. The MAD is probably the most common robust estimator of scale in advanced use. If a large random sample is drawn from a normal distribution with variance σ^2, we expect $\mathrm{E}\{\mathrm{MAD}(\mathbf{X})\} = 0.6745\sigma$. The MAD is often used to give an auxiliary estimate of scale for other more complicated scale estimators, and for n large, is commonly scaled so that it is asymptotically unbiased for the standard deviation σ for normal data (regardless of the actual distribution of the data).

Rousseeuw & Croux (1993) present two estimators as alternatives to the MAD, commonly referred to as S_n and Q_n. These estimators, like the MAD, trimmed mean, interquartile range, and various others, but unlike A-estimators, provide scale estimates that do not rely on any auxiliary scale estimates. Rousseeuw & Croux's (1993) estimators are proportional to those given in the following definitions.

Definition 2.14 ($\mathbf{S_n}$) S_n *for the observations* $\mathbf{X} = (X_1, \ldots, X_n)$ *is given by*

$$S_n = median_i \left\{ median_j \left| X_i - X_j \right| \right\} . \tag{2.19}$$

Thus, S_n is the median of the median interpoint distances for each observation and is motivated as an analogue to Gini's mean difference with averages replaced by medians.

Definition 2.15 ($\mathbf{Q_n}$) Q_n *for the observations* $\mathbf{X} = (X_1, \ldots, X_n)$ *is given by*

$$Q_n = \left\{ |X_i - X_j| \, ; i < j \right\}_{(k)} \tag{2.20}$$

which is the k*th largest of the* $|X_i - X_j|$ *for* $i < j$*, where* $k = {}^hC_2$ *and* $h = [n/2] + 1$.

Thus Q_n is the kth order statistic of the ${}^nC_2$ interpoint distances. Since h is approximately half the number of observations, k is approximately 0.25 times the number of interpoint distances ${}^nC_2$, and hence Q_n is approximately the lower quartile of the interpoint distances. In the case where $n = 20$, $k = 55$ with ${}^nC_2 = 190$, and hence k is the $100(\frac{55}{190}) = 28.9$th percentile of the ordered interpoint distances.

Typically, leading coefficients (1.1926 for S_n, and 2.2219 for Q_n) are included to achieve asymptotic unbiasedness for the standard deviation of Gaussian data; however they are omitted here. Rousseeuw & Croux (1993) perform a small simulation limited to MAD, S_n, Q_n and the sample standard deviation and show that both S_n and Q_n outperform the MAD for both Gaussian data and Cauchy data. On this basis, these estimators are included in this study.

Note that we consider the size of k in Q_n as fixed, although in general this is a parameter that could be manipulated to optimise the performance of Q_n. Rousseeuw & Croux (1993) give no explicit reason for this particular choice of k. However, they do state that this choice attains the 50% breakdown point of the MAD. (For definition and discussion of breakdown points, see Hoaglin et al. (2000) or Rousseeuw & Croux (1993).)

2.3.2 *A*-estimators of scale

While the sample standard deviation, Gini's mean difference, the interquartile range, the MAD and S_n do not have any associated parameters, the A-estimators of scale are a class of estimators which, like the trimmed standard deviation and Q_n, do have associated parameters (note that we choose not to manipulate the parameter k for Q_n in this study). In fact there are many different opportunities to tune the A-estimators through parameter choice, and choice of weighting function. In Lax's study, as here, certain parameter values are chosen and consequently the analysis focuses on the joint hypothesis that the class of estimator with that particular parameter is the "best" robust scale estimator. Much more analysis would be required to find the best class/parameter combination.

Before introducing the A-estimator, it is necessary to define and motivate the M-estimator of location. M-estimators are commonly used to give robust location estimates, and empirical studies have shown that these perform extremely well in a variety of circumstances (see Hoaglin et al. (2000) and the references therein). We motivate the form of the M-estimators through the following example for the location parameter of the normal distribution.

Estimation of the location parameter μ for a random sample $\mathbf{X}$ from the $\mathcal{N}(\mu, \sigma^2)$ distribution using maximum likelihood requires maximisation of the likelihood function

$$L(\mu; \mathbf{X}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(X_i-\mu)^2}{\sigma^2}}$$

where we assume σ is known. The log-likelihood can be written

$$\ln L(\mu; \mathbf{X}, \sigma) = -\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

and thus maximisation of the likelihood is equivalent to maximisation of the log-likelihood, which in turn is equivalent to *minimisation* of the function

$$Q(\mu; \mathbf{X}, \sigma) = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2. \tag{2.21}$$

Differentiating $Q(\mu; \mathbf{X}, \sigma)$ with respect to μ, we solve

$$\sum_{i=1}^{n}(X_i - \mu) = 0$$

which of course yields the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. This process can be extended to motivate M-estimators of location, for which we replace the quadratic loss function $\rho(u) = u^2$ in (2.21) by a general function, symmetric about $u = 0$, and increasing in $|u|$.

An M-estimator T_n is the choice of T which minimises the objective function

$$\sum_{i=1}^{n}\rho\left(\frac{X_i - T}{cS_0}\right) \tag{2.22}$$

where S_0 is an auxiliary estimate of scale (typically MAD), c is a positive constant, and $\rho(u)$ is an even function. Differentiating with respect to T, an alternative specification of the M-estimator is that it is the solution to the equation

$$\sum_{i=1}^{n}\psi\left(\frac{X_i - T_n}{cS_0}\right) = 0 \tag{2.23}$$

where $\psi(u) = \frac{d}{du}\rho(u)$ is an odd function. The most popular choice of $\psi(u)$ is Tukey's biweight ψ-function, which is

$$\psi(u) = \begin{cases} u(1-u^2)^2 & |u| \leq 1 \\ 0 & |u| > 1. \end{cases} \tag{2.24}$$

A table of common ψ-functions and their associated ρ-functions is given in Hoaglin et al. (2000). Goodall (2000) provides the general link between a ψ-function and a target density function. As seen above $\psi(u) \propto u$ is consistent with underlying Gaussianity, and in general a random sample from a distribution with density

$$f(x; \mu) \propto \exp\left[-\int_{\mu}^{x}\psi(v; \mu)dv\right]$$

yields the maximum likelihood estimate T_n where T_n is the M-estimator satisfying (2.23). In the Gaussian case, $\psi(x;\mu) = (x-\mu)/\sigma$ and

$$f(x;\mu) \propto e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

as required. Goodall (2000) points out that ψ-functions like the biweight (2.24) that cut off at a finite u, do not have any associated target distribution.

Substituting $\psi(u) = u\,w(u)$ in (2.23) we find

$$\sum_{i=1}^{n} \left(\frac{X_i - T_n}{cS_0}\right) w\left(\frac{X_i - T_n}{cS_0}\right) = 0$$

which can be rearranged to give

$$T_n = \frac{\sum_{i=1}^{n} w\left(\frac{X_i - T_n}{cS_0}\right) X_i}{\sum_{i=1}^{n} w\left(\frac{X_i - T_n}{cS_0}\right)}. \tag{2.25}$$

Since $\psi(u)$ is an odd function, $w(u)$ is an even function (i.e., symmetric about $u = 0$), and (2.25) gives T_n as a weighted average of the standardised scores $U_i = (X_i - T_n)/cS_0$. The biweight ψ-function has corresponding weight function

$$w(u) = \begin{cases} (1-u^2)^2 & |u| \leq 1 \\ 0 & |u| > 1 \end{cases} \tag{2.26}$$

and this is used to downweight the observations in the location estimate (2.25). (2.26) is the biweight function used by `loess` to downweight observations (see Appendix A), in a process related to the iterated solution of (2.25).

We are now in a position to define the A-estimator of scale, which is the finite sample equivalent of the asymptotic variance of an M-estimator and depends on the choice of ψ-function, and the underlying distribution of the data. The A-estimator is derived in Huber (1981) using techniques not discussed here; however an alternative derivation follows.

It is well known that the sample mean of a random sample has variance σ^2/n, and hence

$$\sqrt{n\text{var}(\bar{X})} = \sigma.$$

In order to estimate σ, we might consider estimating the sample variance of realisations of $\bar{X}$. By analogy to this relationship, Lax motivates the use of the asymptotic variance of an M-estimator to obtain an estimator of scale.

Theorem 2.6 *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be independent and identically distributed symmetric random variables from the location-scale family with location parameter* μ_0 *and scale parameter* σ. *Then the* M*-estimator* T_n *of* μ_0 *has an asymptotic normal distribution with mean* μ_0 *and variance given via*

$$\lim_{n\to\infty} var\{\sqrt{n}(T_n - \mu_0)\} = \sigma^2 \frac{E\{\psi(U)^2\}}{E\{\psi'(U)\}^2} \tag{2.27}$$

where $U \sim U_i = (X_i - \mu_0)/\sigma$ *is the standardised score,* $\psi(u) = \frac{d}{du}\rho(u)$ *is an odd function as specified in (2.23), and* $\rho(u)$ *is a twice-differentiable even function.*

Proof We have the observations $\mathbf{X} = (X_1, \ldots, X_n)$ and these are independent and identically distributed symmetric random variables with location μ_0 and scale σ. Define

$$Q_n(\mu) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{X_i - \mu}{\sigma}\right)$$

where $\rho(u)$ is an even function, μ is a parameter and σ is assumed known. Differentiating with respect to μ, we obtain

$$Q_n'(\mu) = -\frac{1}{\sigma n}\sum_{i=1}^{n} \psi\left(\frac{X_i - \mu}{\sigma}\right)$$

where $\psi(u) = \frac{d}{du}\rho(u)$ is an odd function. By definition, the M-estimator T_n minimises $Q_n(\mu)$ and satisfies $Q_n'(T_n) = 0$. Taking a Taylor Series expansion of $Q_n'(\mu)$ about T_n

$$\begin{aligned} -\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n} \psi\left(\frac{X_i - \mu}{\sigma}\right) &= \sqrt{n}Q_n'(\mu) = \sqrt{n}Q_n'(T_n) + \sqrt{n}(T_n - \mu)Q_n''(\bar{\mu}) \\ &= 0 - \sqrt{n}(T_n - \mu)\frac{1}{\sigma^2 n}\sum_{i=1}^{n} \psi'\left(\frac{X_i - \bar{\mu}}{\sigma}\right) \end{aligned}$$

where $\bar{\mu}$ is a random variable such that $|T_n - \bar{\mu}| < |T_n - \mu_0|$. Evaluating at $\mu = \mu_0$, and taking the limit as $n \to \infty$, we obtain

$$\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n} \psi\left(\frac{X_i - \mu}{\sigma}\right)\Bigg|_{\mu=\mu_0} \longrightarrow \sqrt{n}(T_n - \mu_0)\frac{1}{\sigma^2}\mathrm{E}\{\psi'(U)\}$$

where $U \sim U_i = (X_i - \mu_0)/\sigma$ is the standardised score with zero location and unit scale. T_n converges to to μ_0 as n increases (Huber 1981, Corollary 2.2) and it follows that $\bar{\mu}$ also converges to μ_0 and the right hand side is as stated. Hence, in the limit,

$$\sqrt{n}(T_n - \mu_0) = \sqrt{n}\frac{\frac{1}{\sigma n}\sum\psi(U_i)}{\frac{1}{\sigma^2}\mathrm{E}\{\psi'(U)\}} = \frac{\sigma}{\sqrt{n}}\frac{\sum\psi(U_i)}{\mathrm{E}\{\psi'(U)\}}.$$

Now, define $\epsilon_i = \psi(U_i)$. Since the X_i are independent and identically distributed, the ϵ_i are too, with mean $\mathrm{E}(\epsilon_i) = \mathrm{E}\{\psi(U_i)\} = 0$, since $\psi(u)$ is an odd function and U_i is symmetrically distributed, and variance $\mathrm{var}(\epsilon_i) = \mathrm{E}\{\psi(U_i)^2\}$. In addition, for large n, application of the Central Limit Theorem yields

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \sim \mathcal{N}(0, \mathrm{E}\{\psi(U)^2\}).$$

Thus, as $n \to \infty$

$$\begin{aligned} \sqrt{n}(T_n - \mu_0) = \frac{\sigma}{\sqrt{n}} \frac{\sum \psi(U_i)}{\mathrm{E}\{\psi'(U)\}} &\sim \frac{\sigma}{\mathrm{E}\{\psi'(U)\}} \mathcal{N}\left(0, \mathrm{E}\{\psi(U)^2\}\right) \\ &\sim \mathcal{N}\left(0, \sigma^2 \frac{\mathrm{E}\{\psi(U)^2\}}{\mathrm{E}\{\psi'(U)\}^2}\right) \end{aligned}$$

as required. □

We note that the variance depends not only on the choice of $\psi(u)$, but also on the underlying distribution of X. Choice of the biweight ψ-function in (2.24) ensures that the required moments exist even for long-tailed distributions like the slash.

In the case of Gaussian X_i, as shown above, the form of the likelihood function suggests $\rho(x) \propto x^2$. For this choice of $\rho(u)$, it follows without loss of generality that $\psi(x) = x$ and $\psi'(x) = 1$. The M-estimator is $T_n = \bar{X}$ and thus

$$\sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, \sigma^2)$$

as required.

The A-estimator is defined to be the finite sample equivalent of the asymptotic variance of the M-estimator as follows.

Definition 2.16 (A-estimator) *The A-estimator for the observations $\mathbf{X} = (X_1, \ldots, X_n)$, with ψ-function $\psi(u) = uw(u)$ is*

$$s_\psi(\mathbf{X}; c, S_0) = \left[\frac{1}{n-1} \frac{\sum_{i=1}^{n} w(U_i)^2 (X_i - M)^2}{\left[\frac{1}{n} \sum_{i=1}^{n} \psi'(U_i)\right]^2} \right]^{\frac{1}{2}} \tag{2.28}$$

where $U_i = (X_i - M)/cS_0$, M is an auxiliary estimator of location, S_0 is an auxiliary estimator of scale, c is a positive scaling constant, and $\psi'(u) = \frac{d}{du}\psi(u)$.

It is clear from Definition 2.16 that the A-estimator is the finite sample equivalent of (2.27), and is a weighted average of the squared deviations, with weights determined

by the ratio of $w(U_i)^2$ and the denominator. For a given collection of data, the denominator is a positive scale factor; however it will differ across collections.

Unlike Lax, in order to limit the number of distinct estimates possible, we consider only a single ψ-function: the biweight. Thus $\psi(u)$ and $w(u)$ are defined in (2.24) and (2.26), and we investigate appropriate combinations of S_0 and c. As with many other estimators requiring an auxiliary location estimator, we choose M to be the sample median.

2.3.3 Maximum likelihood estimator for the t-distribution

The robust estimators we have considered so far have been designed to mitigate the effect of extreme observations by considering order statistics (e.g. MAD, IQR, S_n, Q_n) or by taking weighted averages of the "well behaved" observations (e.g. the trimmed standard deviations and A-estimators). Robustness has been achieved by tuning the estimators by choice of which order statistic to use or to use weighting schemes which give weights decreasing in the size of the observations. Evaluation of our choice is by simulation with the performance criterion being triefficiency. As an alternative, we propose the family of t-distributions as an intermediate distribution (one that might successfully model the "goodness" of the Gaussian distribution, but also reflect heavy tailed behaviour) and optimise the scale estimate for this target distribution.

Definition 2.17 (t_ν random variable) *The t_ν random variable with location parameter μ, scale parameter σ and ν degrees of freedom follows the Gaussian compound scale model of Definition 2.9 with S_i a chi-squared random variable, with ν degrees of freedom, divided by ν.*

When $\nu \to \infty$ the t_ν distribution is equivalent to the $\mathcal{N}(\mu, \sigma^2)$, and for $\nu = 1$, the t_ν distribution is the Cauchy, with mean, variance and all higher order moments infinite. In general, the kth central moment for the t_ν distribution is defined if $k < \nu$. As ν increases, statistics measuring the heaviness of the tails such as kurtosis, where defined, and the tail-weight index (both given in Table 2.1 for selected ν) decrease monotonically in ν. Thus we might expect the t-distribution to be sufficiently flexible for modelling the compromise distribution. We hope that the ML estimator for the target t_ν distribution will also perform well under the triefficiency criterion. As with

the A-estimator and its tuning constant, we will rely on simulation to identify the appropriate choice of ν.

Since the t_ν random variable follows a Gaussian compound scale model, we can apply Theorems 2.1 and 2.2 to obtain the maximum likelihood recursions for the unknown parameters μ and σ.

Theorem 2.7 *Consider a random sample* $\mathbf{X} = (X_1, \ldots, X_n)$ *from a t-distribution with location* μ, *scale* σ *and* ν *degrees of freedom, where* μ *and* σ *are unknown and* ν *is known. The maximum likelihood estimators of* μ *and* σ *are found by iterating equations (2.8) and (2.9) with*

$$E_0(S_i|\mathbf{X}) = \frac{\nu+1}{\nu}\left(1 + \frac{(X_i - \hat{\mu}_0)^2}{\nu\hat{\sigma}_0^2}\right)^{-1}$$

where $\hat{\mu}_0$ *and* $\hat{\sigma}_0^2$ *are the previous estimates of* μ *and* σ^2 *respectively.*

Proof The Student's t-distribution with ν degrees of freedom corresponds to the case where S_i is an independent χ^2_ν random variable divided by ν, and hence the conditions of both Theorems 2.1 and 2.2 are met. In particular

$$\mathrm{E}_0(S_i|\mathbf{X}) = G\left(\frac{1}{2}\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2\right)$$

where $G(t) = -d\ln M(t)/dt$ and $M(t)$ is the Laplace transform

$$M(t) = \int_0^\infty e^{-ts}\sqrt{s}dF_{S_i}(s) \qquad (t \geq 0).$$

Since S_i has cdf $P(S_i < s) = P(\chi^2_\nu < s\nu)$, its density is given by

$$f_{S_i}(s) = \nu f_{\chi^2_\nu}(\nu s) = \frac{\nu}{\Gamma(\frac{\nu}{2})}\left(\frac{1}{2}\right)^{\frac{\nu}{2}}(\nu s)^{\frac{\nu}{2}-1}e^{-\frac{1}{2}\nu s} \qquad (s > 0)$$

for all i. Thus

$$\sqrt{s}f_{S_i}(s) \propto s^{\frac{1}{2}(\nu+1)-1}e^{-\frac{1}{2}\nu s}$$

which is in turn proportional to a gamma density function with parameters $\frac{1}{2}(\nu+1)$ and $\frac{1}{2}\nu$, and it follows that $M(t)$ is the moment generating function of such a gamma random variable, i.e.

$$M(t) = \left(\frac{\nu}{\nu+2t}\right)^{\frac{1}{2}(\nu+1)} \qquad (t > -\tfrac{1}{2}\nu).$$

Hence

$$G(t) = \frac{\nu+1}{\nu}\left(1+\frac{2t}{\nu}\right)^{-1}$$

and applying Theorem 2.2,

$$\mathrm{E}_0(S_i|\mathbf{X}) = \frac{\nu+1}{\nu}\left(1+\frac{(X_t-\hat{\mu}_0)^2}{\nu\hat{\sigma}_0^2}\right)^{-1}$$

as required. □

Application of Theorems 2.1 and 2.7 for sample data yields the maximum likelihood estimates of μ and σ^2 under the assumption that the data is a random sample from the t_ν distribution. We use Theorem 2.7 to motivate two different scale estimators for examination in the simulation. The first of these is the fully iterated ML estimator for selected ν as defined in Theorem 2.7. While not optimal for any of the three corners, a carefully chosen ν might allow triefficient estimation of scale.

The second estimator motivated by Theorem 2.7 is obtained by specifying initial estimates for the EM algorithm, $\hat{\mu}_0 = M$ and $\hat{\sigma}_0 = cS_0$ for some positive constant c, and performing only a single iteration. Without updating the location estimate, we obtain an estimator of the following form.

Definition 2.18 (One-step t-estimator) *The one-step t-estimator of scale for the observations $\mathbf{X} = (X_1, \ldots, X_n)$ is given by*

$$s_t(\mathbf{X}; c, S_0) = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{1+U_i^2}\right)(X_i-M)^2\right]^{\frac{1}{2}} \tag{2.29}$$

where $U_i = (X_i - M)/cS_0$, M is a an auxiliary estimate of location, S_0 is an auxiliary estimate of scale, and c is a positive scaling constant.

This estimator is similar in form to an A-estimator; however in this case the weight function is very simple. If S_0 is a consistent estimator of scale for a t-distribution with ν degrees of freedom, then $c = \sqrt{\nu}$ is required, as is the multiplicative constant $\frac{\nu+1}{\nu}$ to make $s_t(\mathbf{X}; c, S_0)$ a consistent estimator of σ. If iterated to convergence for a random sample from the t_ν distribution (as in Theorem 2.7) we know the optimal downweighting of the observations is attained using the weight chosen in (2.29); however this weight function may also be useful for data that doesn't follow a t-distribution. Thus, as for the A-estimator, we select S_0 and search for the scaling constant $c > 0$ which allows the greatest triefficiency.

2.4 Methodology

In the following section, we discuss the design of the simulation.

2.4.1 Evaluation criteria

Rather than calculating the efficiency of an estimator in the usual way, by taking the ratio of the minimum attainable variance and that estimator's variance, the performance of scale estimators is usually assessed using the variance of log estimates. Thus the efficiency of a scale estimator $S_1(\mathbf{X})$ relative to another estimator $S_2(\mathbf{X})$ is

$$\text{eff}(S_1, S_2) = \frac{\text{var}[\ln S_2(\mathbf{X})]}{\text{var}[\ln S_1(\mathbf{X})]}. \tag{2.30}$$

If S_1 is less efficient than S_2, its variance will be higher, and so $\text{eff}(S_1, S_2) < 100\%$.

The first benefit of using the log transformation is that any constant multipliers in the scale estimators disappear, since for constant b,

$$\text{var}[\ln\{bS(\mathbf{X})\}] = \text{var}[\ln S(\mathbf{X})].$$

Hence, proportional biases in the estimators are not important and we do not need to select constants for asymptotic (or finite sample) unbiasedness. The second advantage is that the distributions of the scale estimates themselves are made more symmetrical by the log transform due to the (generally longer) upper tail being condensed relative to the lower one. This symmetrising effect makes use of the sample variance to estimate the theoretical variances in (2.30) more suitable, particularly for long-tailed distributions like the slash.

Definition 2.19 (Sample efficiency of a scale estimator) *The sample efficiency of a scale estimator $S(\mathbf{X})$ can be estimated using m independent realisations of the observations $\mathbf{X} = (X_1, \ldots, X_n)$ and*

$$\widehat{eff}(S) = \frac{\textit{sample variance of } \ln\hat{\sigma}_1, \ldots, \ln\hat{\sigma}_m}{\textit{sample variance of } \ln S(\mathbf{X})_1, \ldots, \ln S(\mathbf{X})_m} \tag{2.31}$$

where $\hat{\sigma}_i$ is the maximum likelihood scale estimate for sample i and $S(\mathbf{X})_i$ is the scale estimate for sample i.

This definition follows from (2.30), where $S_2(\mathbf{X})$ is taken to be the ML estimator, and $S_1(\mathbf{X})$ the estimator of interest. The efficiency is estimated by repeatedly sampling, and estimating the sample variances of the two sets of m scale estimates.

Lax uses the variance of the log estimates to compute efficiencies, as in Definition 2.19, whereas Rousseeuw & Croux (1993) use the standardised variances

$$\frac{m(\text{sample variance of } S(\mathbf{X})_1, \ldots, S(\mathbf{X})_m)}{(\text{sample mean of } S(\mathbf{X})_1, \ldots, S(\mathbf{X})_m)^2} \tag{2.32}$$

where again $S(\mathbf{X})_i$ is the scale estimate for sample i. Efficiencies are calculated as the ratio of the standardised variance for the ML estimator to the standardised variance of the estimator of interest. The standardised variance is m times the squared coefficient of variation, and also is invariant to location and scale transformations of the data, however it does not benefit from the symmetrising effect of the log transformation.

Both evaluation criteria are considered in this study. However, unless otherwise stated, any mention of efficiency relates to that in Definition 2.19, and in particular, all estimators are benchmarked against the sampling variation of the log maximum likelihood estimates. We note that since $\hat{\sigma}_i$ is the maximum likelihood estimate of σ, then $\ln \hat{\sigma}_i$ is also the ML estimate of $\ln \sigma$ and hence use of the ML estimates as the benchmark is particularly attractive.

2.4.2 To swindle, or not to swindle?

One of the first concerns in reproducing Lax's simulation results was the "swindle", or variance reduction technique, used by Lax to increase the effective simulation size, while keeping the actual runs to 1000, 640 and 640 for the normal, one-wild and slash distributions respectively. The swindle is described by Simon (1976) and outlined in more detail below. In the case of the normal distribution, the swindle amounts to standardising the observations using the sample mean and standard deviation, processing the standardised scores, and thus eliminating the variation in the sample statistics. Each estimator's sample variability (actually the variance of the log estimator) is then written as some addition to the known variability of the sample standard deviation for normal data.

As we have seen, each of the three distributions used in the Lax study follows the Gaussian compound scale model of Definition 2.9, and in each case, the observations may be written as

$$X_i = \mu + \sigma \frac{Z_i}{Y_i}$$

where the X_i are the observations, the Z_i are independent, standard normal random variables, and the Y_i are positive random variables, independent of the Z_i, but not

necessarily independent of one another. In the case of the normal distribution, $Y_i = 1$ for all i (clearly a dependent sequence), and for the slash distribution we require Y_i to be independent uniform random variables on the interval $[0, 1]$. It follows from the proof to Theorem 2.5, that for a one-wild sample of size n, we require a randomly selected Y_i to be equal to $\frac{1}{10}$ and the remaining Y_i to be one. This is a non-trivial example of a dependent sequence.

Given the data $\mathbf{X} = (X_1, \ldots, X_n)$ and conditioning on $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the Generalised Least Squares (GLS) estimators of μ and σ^2 are obtained by minimising the weighted sum of squared errors

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{Y_i^2 (X_i - \mu)^2}{\sigma^2}$$

with respect to $\boldsymbol{\theta} = (\mu, \sigma^2)$, giving

$$\hat{\mu} = \frac{\sum_{i=1}^{n} Y_i^2 X_i}{\sum_{i=1}^{n} Y_i^2} \tag{2.33}$$

which is a weighted average of the observations $\mathbf{X}$, and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} Y_i^2 (X_i - \hat{\mu})^2}{n - 1} \tag{2.34}$$

which is a weighted average of the squared deviations of the X_i about $\hat{\mu}$. Since, given $\mathbf{Y}$, these are GLS estimators, they are the Ordinary Least Squares (OLS) estimators for a transformed model (Seber 1977), and these latter estimators are independent, complete and sufficient statistics for μ and σ^2 (Graybill 1976, Theorem 6.2.1). In particular, we can conclude that given $\mathbf{Y}$, $\hat{\sigma}$ is a complete, sufficient statistic for σ.

In an operation related to standardisation, we form the configuration vector $\mathbf{C} = (C_1, \ldots, C_n)$ with elements

$$C_i = \frac{X_i - \hat{\mu}}{\hat{\sigma}}.$$

In the case where the X_i are normal with $Y_i = 1$ for all i, $\hat{\mu}$ is the sample mean, $\hat{\sigma}$ is the sample standard deviation, and the C_i are identically the standardised scores.

Interpreting $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z} = (Z_1, \ldots, Z_n)$, and $\mathbf{C}$ as column vectors, we can write in general,

$$\mathbf{C} = \frac{\mathbf{X} - \hat{\mu}\mathbf{1}}{\hat{\sigma}}$$

where $\mathbf{1}$ is the unit column vector of length n, and where

$$X_i - \hat{\mu} = \sigma \left(\frac{Z_i}{Y_i} - \frac{\sum_{i=1}^{n} Y_i Z_i}{\sum_{i=1}^{n} Y_i^2} \right)$$

by the definitions of X_i and $\hat{\mu}$. This yields

$$(\mathbf{X} - \hat{\mu}\mathbf{1}) = \sigma \left(\mathbf{D}^{-1} - \frac{\mathbf{1}\mathbf{Y}^\top}{\mathbf{Y}^\top \mathbf{Y}} \right) \mathbf{Z}$$

where $\mathbf{D} = \text{diag}(Y_i)$, and this is clearly independent of μ. Moreover,

$$\hat{\sigma}^2 = \frac{\sigma^2}{n-1} \sum_{i=1}^{n} Y_i^2 (X_i - \hat{\mu})^2 = \frac{\sigma^2}{n-1} \left[\mathbf{Z}^\top \left(\mathbf{I} - \frac{\mathbf{Y}\mathbf{Y}^\top}{\mathbf{Y}^\top \mathbf{Y}} \right) \mathbf{Z} \right]$$

which is also independent of μ. Dividing $(\mathbf{X} - \hat{\mu}\mathbf{1})$ by $\hat{\sigma}$, the σ terms cancel, and we see that $\mathbf{C}$ is a function of $\mathbf{Z}$, which is a vector of independent standard normals, and the Y_i, which we are conditioning on, alone. Thus, given $\mathbf{Y}$, the distribution of $\mathbf{C}$ does not depend on μ or σ and so $\mathbf{C}$ is an ancillary statistic. In addition, since $\hat{\sigma}$ is a complete, sufficient statistic for σ, Basu's Theorem (Lehmann & Casella 1998, Theorem 1.6.21) states that given $\mathbf{Y}$, the random vector $\mathbf{C}$ and the scale estimator $\hat{\sigma}$ are independent. This result is also stated by Simon (1976).

Now we derive the swindle. Due to the principle property of a scale estimator given in (2.4), and since $\hat{\sigma} > 0$, the scale estimator $S(\mathbf{X})$ satisfies

$$S(\mathbf{X}) = \hat{\sigma} S(\mathbf{C}). \tag{2.35}$$

So the scale estimator of the observations $\mathbf{X}$ is simply $\hat{\sigma}$ times the scale estimator of the configuration vector $\mathbf{C}$. We wish to calculate the efficiency of the scale estimator using (2.31) in the special case where $\mu = 0$ and $\sigma^2 = 1$, and hence we need to estimate $\text{var}\{\ln S(\mathbf{X})\}$. The most straight-forward method of doing this is to simulate many realisations of $\ln S(\mathbf{X})$, and estimating their (population) variance using the sample variance of the realised values. Thus, we estimate

$$\widehat{\text{var}}\{ \ln S(\mathbf{X}) \} = \text{sample variance of } \ln S(\mathbf{X})_1, \ldots, \ln S(\mathbf{X})_m \tag{2.36}$$

where m samples $\mathbf{X}$ are simulated, and $\widehat{\text{var}}$ denotes the sample variance. We then use this quantity in (2.31) to obtain an efficiency estimate. This is described as straight Monte Carlo sampling in the discussion below.

Noting the relationship (2.35), the variance on the left hand side of (2.36) admits the following decomposition

$$\widehat{\text{var}}\{ \ln S(\mathbf{X}) \} = \widehat{\text{var}}(\ln \hat{\sigma}) + \widehat{\text{var}}\{ \ln S(\mathbf{C}) \} + 2\, \widehat{\text{cov}}\{ \ln \hat{\sigma}, \ln S(\mathbf{C}) \} \tag{2.37}$$

where $\widehat{\text{var}}$ and $\widehat{\text{cov}}$ are the sample variance and covariance respectively. The decomposition (2.37) shows the three sources of variation in the statistic we compute using (2.36).

An alternative method of estimating $\text{var}\{\ln S(\mathbf{X})\}$, and the basis of the swindle used by Lax (1985), is obtained by conditioning on $\mathbf{Y}$, which is of course available due to the generation of $\mathbf{X}$. Conditioning on $\mathbf{Y}$, the variance of interest can be decomposed as follows:

$$\begin{aligned}\text{var}\{\ln S(\mathbf{X})\} &= \text{var}(\ln\hat{\sigma} + \ln S(\mathbf{C})) \\ &= \text{var}\{\text{E}(\ln\hat{\sigma} + \ln S(\mathbf{C})|\mathbf{Y})\} + \text{E}\{\text{var}(\ln\hat{\sigma} + \ln S(\mathbf{C})|\mathbf{Y})\} \qquad (2.38)\end{aligned}$$

since in general $\text{var}(X) = \text{var}\{\text{E}(X|Y)\} + \text{E}\{\text{var}(X|Y)\}$. Noting that $\sigma = 1$, given $\mathbf{Y}$, $(n-1)\hat{\sigma}^2$ is chi-squared with $n-1$ degrees of freedom, $\text{E}(\ln\hat{\sigma}|\mathbf{Y})$ is a fixed constant, and hence

$$\text{var}\{\text{E}(\ln\hat{\sigma} + \ln S(\mathbf{C})|\mathbf{Y})\} = \text{var}\{\text{E}(\ln S(\mathbf{C})|\mathbf{Y})\}. \qquad (2.39)$$

In addition, given $\mathbf{Y}$, $\hat{\sigma}$ and $\mathbf{C}$ are independent, then

$$\text{E}\{\text{var}(\ln\hat{\sigma} + \ln S(\mathbf{C})|\mathbf{Y})\} = \text{E}\{\text{var}(\ln\hat{\sigma}|\mathbf{Y})\} + \text{E}\{\text{var}(\ln S(\mathbf{C})|\mathbf{Y})\}. \qquad (2.40)$$

Substituting (2.39) and (2.40) in (2.38), and noting

$$\text{var}\{\text{E}(\ln S(\mathbf{C})|\mathbf{Y})\} + \text{E}\{\text{var}(\ln S(\mathbf{C})|\mathbf{Y})\} = \text{var}(\ln S(\mathbf{C}))$$

we have

$$\text{var}\{\ln S(\mathbf{X})\} = \text{var}\{\ln S(\mathbf{C})\} + \text{var}\left(\ln\sqrt{\chi^2_{n-1}/(n-1)}\right) \qquad (2.41)$$

where $\chi^2_{n-1} = (n-1)\hat{\sigma}^2$ is a chi-squared random variable with $n-1$ degrees of freedom.

An approximation to the variance of the log of a chi-squared random variable divided by its degrees of freedom is given in Abramowitz & Stegun (1968) as

$$\text{var}\left\{\ln\left(\chi^2_\nu/\nu\right)\right\} = \frac{2}{\nu-1}\left(1 - \frac{1}{3(\nu-1)^2}\right) + O((\nu-1)^5)$$

and hence

$$\text{var}\left(\ln\sqrt{\chi^2_{n-1}/(n-1)}\right) \approx \frac{1}{2(n-2)}\left(1 - \frac{1}{3(n-2)^2}\right)$$

which yields approximately 0.027749 when $n = 20$. Thus, using the swindle, we can estimate the variance of interest using

$$\begin{aligned}\widehat{\text{var}}\{\ln S(\mathbf{X})\} &= 0.027749 + \widehat{\text{var}}\{\ln S(\mathbf{C})\} \\ &= 0.027749 + \text{sample variance of } \ln S(\mathbf{C})_1, \ldots, \ln S(\mathbf{C})_m \qquad (2.42)\end{aligned}$$

where again, m samples $\mathbf{X}$ are simulated, each of this yielding a configuration vector $\mathbf{C}$.

In the case of Gaussian $\mathbf{X}$, $\hat{\mu}$ is equal to the sample mean and $\hat{\sigma}$ the sample standard deviation, and consequently the configuration vector contains standardised scores. When the scale estimator is the sample standard deviation, $S(\mathbf{C}) = 1$ by definition, and so $\ln S(\mathbf{C}) = 0$ for any observations $\mathbf{X}$. The equality (2.41) above holds, since the sample standard deviation on the left hand side equals $\hat{\sigma}$, and thus has the same distributional properties.

Comparing (2.36) and (2.37) to (2.42), we see that when $\operatorname{var}\{\ln S(\mathbf{C})\}$ is large, the swindle will be of little use. However, when this amount is small relative to the constant 0.027749 (as in the standard deviation for normal $\mathbf{X}$), the swindle will have a large impact on the precision of the estimated variance due to successful elimination of variability of the sample estimates $\widehat{\operatorname{var}}(\ln \hat{\sigma}) + 2\ \widehat{\operatorname{cov}}\{\ln \hat{\sigma}, \ln S(\mathbf{C})\}$.

Whilst undoubtedly the swindle was very important for Lax, it seems that its benefit is limited now. Computing power has increased to the extent that it is much easier to increase precision by increasing the simulation sizes than to allocate the additional computation required to compute the configuration vector for each sample. The swindle is employed to yield an efficiency based on (2.42), (2.31) and 1000 independent samples. This is repeated 100 times to yield 100 efficiency estimates, and these are compared to those obtained through straight Monte Carlo simulation. We find that for some estimators and the normal distribution, the swindle achieves the precision of up to 4000 Monte Carlo samples. However, in other instances, and particularly for the slash distribution, the precision attained is certainly no greater than for the same number of Monte Carlo samples. These findings are reflected in the comments above and confirm the remarks of Gross (1976): "the swindle works better for distributions which are 'close' to the Gaussian than for those not so close, and additionally, better for estimators which are relatively 'good' in a situation than those which are not" (Gross 1976, page 411).

Figure 2.7 examines a few estimators more carefully, and allows us to estimate the gain of the swindle in these cases. The first block of the plot shows that efficiencies calculated for the sample standard deviation, and one-wild samples using the swindle and 1000 samples, have a distribution similar to those based on almost 4000 Monte Carlo samples. For the slash distribution and the sample standard deviation, the gains are much smaller, with the efficiencies being almost as variable as those

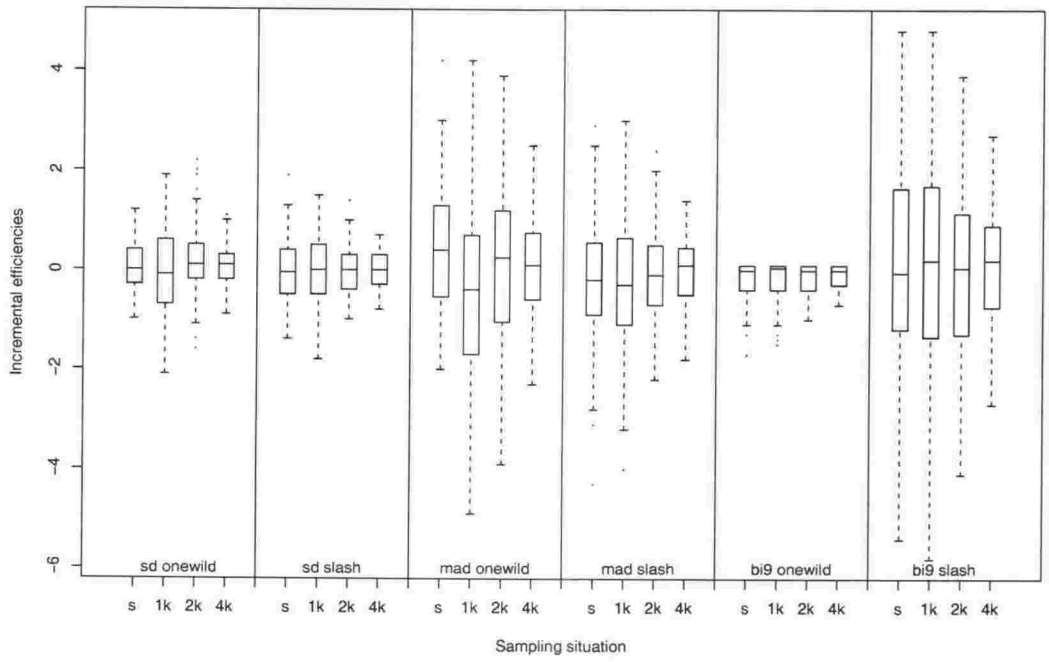

Figure 2.7. Efficiency distributions for various estimators using the swindle outlined in Section 2.4.2. The sampling situations are s=1000 samples using the swindle, 1k, 2k and 4k are 1000, 2000, and 4000 Monte Carlo samples respectively. The estimators, corresponding to the six blocks in the plot, are: standard deviation for the one-wild, standard deviation for the slash, MAD for the one-wild, MAD for the slash, biweight A-estimator with $c = 9$ for the one-wild and biweight A-estimator with MAD and $c = 9$ for the slash. Efficiencies have been translated using the average efficiency from this study for that situation, and are based on 100 repetitions of the respective sampling situations, for samples of size $n = 20$.

First run				*Mth run*	
Normal	One-wild	Slash		Normal	One-wild	Slash
$S_j(\mathbf{X})_1$	$S_j(\mathbf{X}')_1$	$S_j(\mathbf{X}'')_1$	$\cdots$	$S_j(\mathbf{X})_1$	$S_j(\mathbf{X}')_1$	$S_j(\mathbf{X}'')_1$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
$S_j(\mathbf{X})_m$	$S_j(\mathbf{X}')_m$	$S_j(\mathbf{X}'')_m$		$S_j(\mathbf{X})_m$	$S_j(\mathbf{X}')_m$	$S_j(\mathbf{X}'')_m$
$\widehat{\text{var}}_{j,1}^{(1)}$	$\widehat{\text{var}}_{j,1}^{(2)}$	$\widehat{\text{var}}_{j,1}^{(3)}$	$\cdots$	$\widehat{\text{var}}_{j,M}^{(1)}$	$\widehat{\text{var}}_{j,M}^{(2)}$	$\widehat{\text{var}}_{j,M}^{(3)}$
$\text{eff}_{j,1}^{(1)}$	$\text{eff}_{j,1}^{(2)}$	$\text{eff}_{j,1}^{(3)}$		$\text{eff}_{j,M}^{(1)}$	$\text{eff}_{j,M}^{(2)}$	$\text{eff}_{j,M}^{(3)}$

Table 2.5. Schematic representation of the simulation for estimator j. At each run of the $M = 100$ runs, $m = 20000$ samples of size $n = 20$ are generated for each corner: normal $\mathbf{X}$, onewild $\mathbf{X}'$ and slash $\mathbf{X}''$. Scale estimator j is evaluated for each of these samples, and the variance of the log estimates collected for each corner; thus $\widehat{\text{var}}_{j,k}^{(\ell)}$ is obtained for each estimator, run $k = 1, \ldots, M$ and distribution $\ell = 1, 2, 3$. Using (2.31) and the variance estimates for the ML estimator, the efficiencies $\text{eff}_{j,k}^{(\ell)}$ are also obtained for each estimator, run $k = 1, \ldots, M$ and distribution $\ell = 1, 2, 3$.

calculated without the swindle. For the MAD, in blocks three and four of the plot, we see a similar effect: the swindle has considerable benefit in the one-wild situation but only a small effect for slash data. For the biweight A-estimator with MAD and $c = 9$, the distribution obtained for both one-wild and slash using the swindle is barely different to that obtained without it. These, and results for other estimators not shown here, demonstrate that the contribution of the swindle is limited in many cases, and hence it will not be used at all in this study.

2.4.3 This simulation

In this study, all simulations are conducted using the statistical software `R` (Ihaka & Gentleman 1996), Version 1.2.2, installed on a Pentium-III personal computer, running the Red Hat Linux 6.2 operating system. The aim of the simulation is to obtain efficiency figures, calculated using (2.31). Rather than restrict ourselves to a single efficiency estimate for each estimator and distribution combination, as in Lax (1985), we repeat the simulation a number of times, and consequently can comment on the precision of the efficiency estimates.

There are several levels of sampling in the simulation. These are described below, and also represented graphically in Table 2.5.

- The basic unit of the simulation is the sample $\mathbf{X} = (X_1, \ldots, X_n)$ (not a random sample in the one-wild case) on which the scale estimates are based. For each sample, a number of competing scale estimators are evaluated. Primarily $n = 20$, however $n = 10$ and $n = 40$ are also investigated.

- The normal samples are generated randomly and then used as a basis for the one-wild and slash samples.
- The normal sample is copied, and then a randomly selected observation is multiplied by 10 in order to obtain a one-wild sample.
- The normal sample is copied and divided by n independently sampled uniform observations on the interval $[0, 1]$ to obtain a slash sample.

Each sample yields a scale estimate from each of the scale estimators in Table Table 2.6 considered.

- In order to obtain an efficiency estimate, $m = 20000$ samples from each distribution are simulated. Thus, we obtain a collection of scale estimates indexed by distribution and estimator, and the sample variance of the log estimates of each class are calculated, and efficiencies formed according to (2.31).

  A run of 20000 independent samples yields a single efficiency for each estimator at each distribution. (Note that Lax's entire study constitutes $m = 1000$ samples of size $n = 20$.)

- In order to ascertain the precision of the efficiency estimates in each situation, we obtain $M = 100$ estimates of each efficiency, by repeatedly processing m samples of size n, as described above.

The estimators considered are listed in Table 2.6 along with a reference code and a point of definition. Those marked with an asterisk are also simulated for samples of size $n = 10$ and 40. Most estimators have been defined in Section 2.3; however in two instances a reference is given to Lax (1985).

2.5 Results

The benefit of modern computing in a study of this nature is immense. The scope now afforded us due to increased speed and storage capabilities opens opportunities for analysis not possible for Lax. The primary focus of this section is on the results for $n = 20$. These are analysed in detail, and compared to the results of Lax and others where possible. Notable differences between these results and those for the sample sizes $n = 10$ and 40 are reported at the end of the section.

Estimator	Parameters	Code	Definition
sample standard deviation	*	sd (ML for normal)	2.7
Gini's mean difference		gini	2.10
trimmed standard deviation	$(p,r) \in \{(0.1,0.1)^*,(0.2,0.15),(0.2,0.2)\}$	s10, s15, s20	2.11
interquartile range		iqr	2.12
median absolute deviation	*	mad	2.13
S_n	*	Sn	2.14
Q_n	$k = {}^hC_2$ where $h = [n/2]+1^*$	Qn	2.15
modified biweight A-estimator	$c = 6$	mbi	Lax (1985)
modified sine A-estimator	$c = 2.1$	msi	Lax (1985)
biweight A-estimator	$S_0 = \text{MAD}$ and $c \in \{9^*,10^*,11,12,13\}$	bi9, bi10, bi11, bi12, bi13	2.16
	$S_0 = S_n$ and $c \in \{6.5^*,7^*.7.5\}$	bs1, bs2, bs3	
	$S_0 = Q_n$ and $c \in \{10.5^*,11^*,11.5\}$	bq1, bq2, bq3	
iterated t-estimator	$\nu \in \{1,2,3,4,6\}$	t1, t2, t3, t4, t6	Theorem 2.7
one-step t-estimator	$S_0 = \text{MAD}$ and $c \in \{4,4.25,\ldots,5.25\}$	tm1, tm2, tm3, tm4, tm5, tm6	2.18
	$S_0 = S_n$ and $c \in \{2.75,3,3.25\}$	ts1, ts2 ts3	
	$S_0 = Q_n$ and $c \in \{4,4.25^*,4.5\}$	tq1, tq2 tq3	
ML estimator: onewild	*	ML	Theorem 2.5
ML estimator: slash	*	ML	Theorem 2.4

Table 2.6. Estimators considered in the simulation for samples of size $n = 20$, the tuning parameters used, and their point of definition. The thirteen estimators marked with an asterisk are also examined for samples of size $n = 10$ and 40. Simple codes are provided for easy reference in Figures 2.9 to 2.18.

2.5.1 The maximum likelihood estimates

We present summary statistics for the individual maximum likelihood scale estimates for 20000 independent samples from each distribution. These estimates are computed using an iterative algorithm (the EM algorithm), and iterations are terminated when the absolute change in the scale estimate from one iteration to the next is less than 10^{-6}, or when the number of iterations reaches 200. This latter condition is used once in the 20000 samples detailed in this section, and occurs when the estimates alternate between two values slightly further apart than 10^{-6}. Since we use the standard normal distribution as the basis for all three distributions, unbiased (squared) maximum likelihood scale estimates should have unit expected value. The realised bias for the sample standard deviation in the normal case is consistent with theory, and the sample variances for the single run of 20000 have average 1.0000 (4dp). The averages for the scale estimates, and scale estimates squared for the entire simulation are given in Table 2.7. The average standard deviation differs from the theoretical value for the normal distribution, 0.9869 given by (2.14), only in the fourth decimal place.

The distributions of the natural log of the maximum likelihood scale estimates are shown in Figure 2.8 for each of the three corners and for 20000 samples of sizes $n = 10$, 20 and 40, and 200 samples of sizes $n = 160$ and 640. All estimates do

	normal	one-wild	slash
maximum likelihood scale	0.9870	0.9606	0.9960
maximum likelihood scale squared	1.0002	0.9487	1.0890

Table 2.7. Average maximum likelihood estimates of scale and scale squared, over 100 simulations and 20000 samples of size 20. The estimators are the sample standard deviation in the normal case, and given by Theorems 2.5 and 2.4 for the one-wild and slash respectively. Population values of scale and scale squared are unity for all distributions.

appear to have a small downward bias from $\ln(1) = 0$, however this is consistent with Jensen's inequality

$$\mathrm{E}(\ln \hat{\sigma}) = \frac{1}{2}\mathrm{E}(\ln \hat{\sigma}^2) \leq \frac{1}{2} \ln \mathrm{E}(\hat{\sigma}^2) = 0$$

where the final equality holds if $\hat{\sigma}^2$ is an unbiased estimator of $\sigma^2 = 1$. As evident in the plot, the estimates for the normal and one-wild distributions, whose samples are identical but for a single value, are very similar for each sample size. The one-wild estimates have slightly greater range than the sample standard deviations in the normal case. The log transformation has elongated the lower tail relative to the upper for these two corners, however the slash estimates are very symmetric under this transformation reflecting the considerable effect the extremely heavy tails of the slash distribution have on even the optimal scale estimates. Not surprisingly, the estimates for the slash samples have a much greater variability, but are still located close to the theoretical value. In each case, we see that the variance of the log estimates decreases as sample size increases, as we would expect. The plot strongly suggests consistency for the true parameter $\sigma = 1$ for each of the three corners, i.e., as n increases, the bias gets smaller, as does the variance of the log estimates.

2.5.2 Simulation results

In this section, we present the results of the entire simulation based on the variances of the log scale estimates. Where possible, we compare the results from this simulation to previously published results, in particular those of Lax (1985). As there have been many estimators examined, as described in the introductory section, estimators have been divided into single-pass estimators, and "the-rest": the A- and t-estimators. Results are presented separately, with comparison made across the groups where appropriate. Average efficiencies for all the estimators are given in a single table in Appendix C.

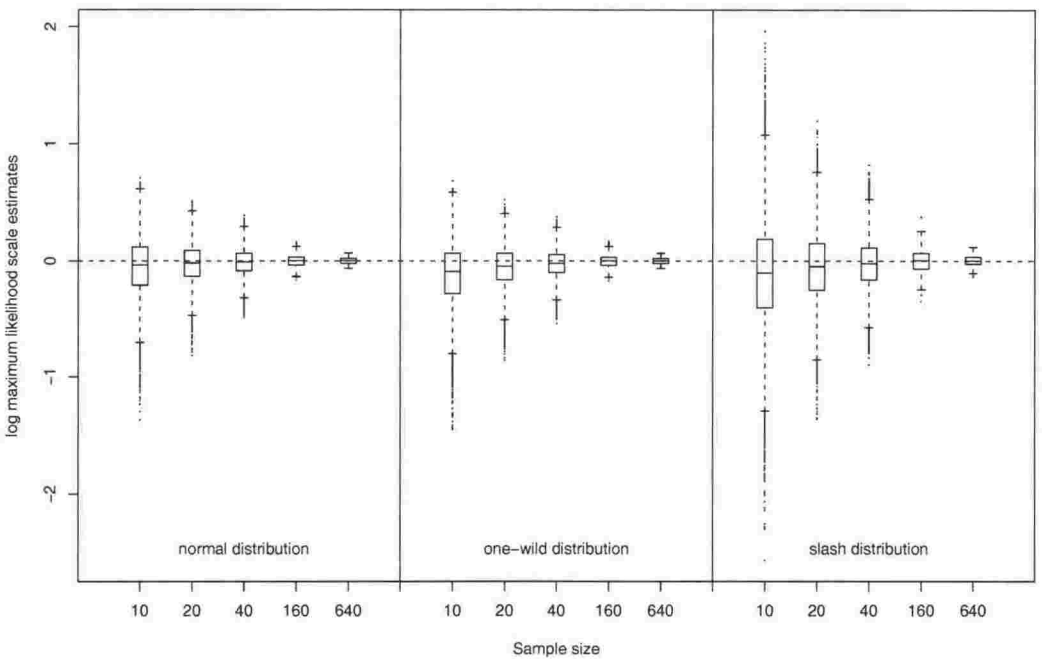

Figure 2.8. Distribution of log maximum likelihood scale estimates for the three corners for 20000 samples of sizes $n = 10$, 20 and 40, and for 200 samples of sizes $n = 160$ and 640. The population value of $\ln \sigma$ is $\ln(1) = 0$ in each case, and this is shown by the horizontal line.

Single-pass estimators

Each simulation run of 20000 samples yields an efficiency estimate for each estimator and each distribution. Efficiencies of the one-pass estimators are compared in Figure 2.9, and average efficiencies are given in Table 2.8. In Figure 2.9, and subsequent figures of this type, median efficiencies for a single corner are connected by line segments, so that we can readily identify performance for that corner. Further, for comparison, all such plots have a y-axis ranging from 0% to 100% efficiency. In addition to the efficiency on the left vertical axis, a non-linear scale measuring the ratio of the standard deviation of the log estimates to the standard deviation of the log ML estimates is provided on the right vertical axis of this plot, and subsequent figures of this type. This ratio is defined

$$\sqrt{\frac{1}{\text{eff}(S)}} = \frac{\text{sample standard deviation of } \ln S(\mathbf{X})_1, \ldots, \ln S(\mathbf{X})_m}{\text{sample standard deviation of } \ln \hat{\sigma}_1, \ldots, \ln \hat{\sigma}_m} \tag{2.43}$$

where $\text{eff}(S)$ is a percentage. This measure gives us a better understanding of the implications of a low efficiency. For example, if a scale estimator has efficiency of 80%, the standard deviation of its log estimates is approximately 1.12 times the standard deviation of the log ML estimates.

The first notable feature of Figure 2.9 is the very poor performance of both the sample standard deviation and Gini's mean difference for one-wild and slash data; Gini's mean difference performing slightly better in each instance. Both are, however, highly efficient for normal data. The next three estimators are the trimmed standard deviations. Various parameter estimates are chosen, and an obvious trade-off between efficiency in the one-wild and slash situations occurs. As we let p and r decrease to the point that a single observation is trimmed, the estimator becomes close to optimal for the one-wild distribution, but hopeless for the slash. In terms of triefficiency, Lax's choice of $p = r = 0.2$ is certainly the best of the three parameter combinations examined here.

The remaining estimators are all robust estimators, and as such their performance for normal data is typically the worst, and their performance for slash data the best. The four estimators: interquartile range (IQR), median absolute deviation (MAD), S_n and Q_n, are all similar in their construction, depending primarily on order statistics. The relative performance of the IQR and MAD is particularly interesting. While the MAD is more efficient for slash data, the IQR is only marginally less efficient in this case, and overall, more triefficient. Under the criterion of triefficiency, we conclude that the IQR is more robust than the MAD, and more suitable generally. This conflicts with the popularity of the MAD in advanced statistical methods (e.g. `loess`). The performance of S_n and Q_n supports their use as alternatives to MAD, and in turn IQR, with particularly high average efficiency in the slash case, of 95.8% and 94.9% respectively. In fact, both MAD and IQR are dominated by S_n and Q_n. We analyse the implications of this for use of MAD as the auxiliary scale estimator of choice in Section 2.5.3.

Table 2.9 offers comparison to Lax's results for sample standard deviation, the trimmed standard deviation with $p = r = 0.2$, and MAD. It also offers comparison to the results of Iglewicz (2000) for the IQR (here comparison is made to results for the fourth spread, which is a close approximation to the IQR). Finally it offers comparison to the results of Rousseeuw & Croux (1993) for S_n and Q_n.

The one-wild efficiency for the sample standard deviation given by Lax appears within sampling error, however the results for the trimmed standard deviation are grossly different. We also see evidence of Lax's failure to benchmark the efficiencies against the minimum variance scale estimators as discussed in Section 2.1.2. Lax reports optimal (100%) efficiency for the trimmed standard deviation at the one-wild

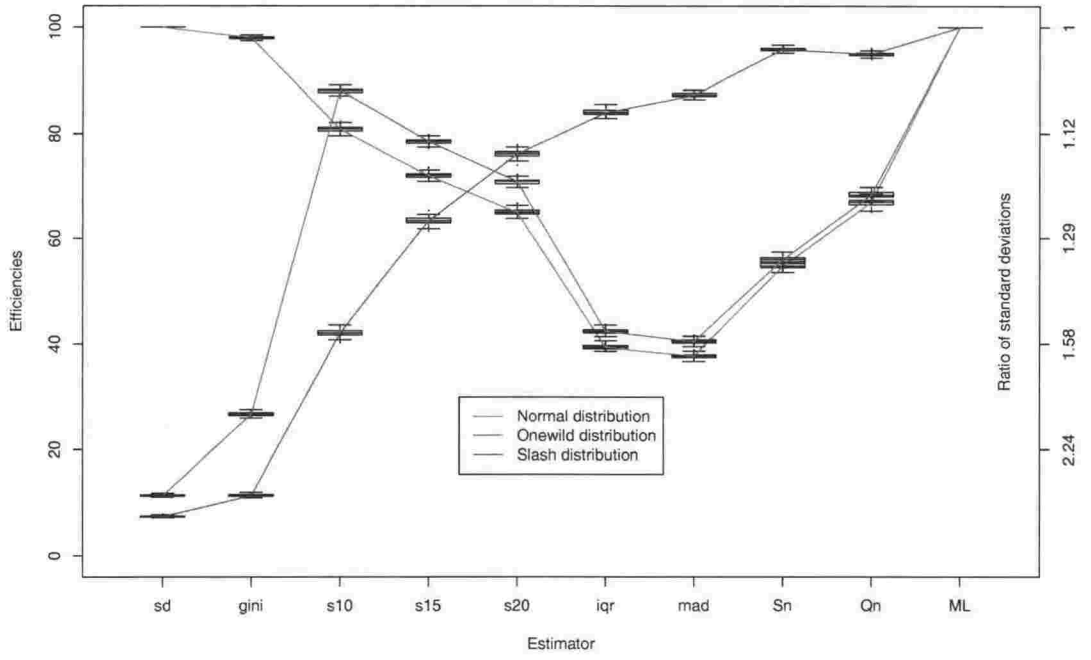

Figure 2.9. Efficiency distributions for the one-pass estimators, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are sd=sample standard deviation, gini=Gini's mean difference, s10=trimmed standard deviation with $p = r = 0.1$, s15=trimmed standard deviation with $p = 0.2$ and $r = 0.15$, s20=trimmed standard deviation with $p = r = 0.2$, iqr=interquartile range, mad=median absolute deviation, Sn=S_n, Qn=Q_n and ML=maximum likelihood. Efficiency is computed using (2.31). The ratio of standard deviations is a non-linear scale given by (2.43).

estimator	normal	one-wild	slash	triefficiency
sample standard deviation	100.0	11.4	7.5	7.5
Gini's mean difference	98.0	26.7	11.4	11.4
trimmed sd with $p = r = 0.2$	65.0	70.8	76.1	65.0
trimmed sd with $p = 0.2$ and $r = 0.15$	72.1	78.6	63.4	63.4
trimmed sd with $p = r = 0.1$	80.9	88.1	42.1	42.1
interquartile range	39.4	42.4	84.0	39.4
median absolute deviation	37.8	40.5	87.3	37.8
S_n	54.7	55.9	95.8	54.7
Q_n	66.9	68.4	94.9	66.9

Table 2.8. Average efficiencies for the one-pass estimators, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31). The triefficiency given is the average over the 100 simulations, rather than the minimum average. The efficiency distributions for these estimators are shown in Figure 2.9.

estimator	normal		one-wild		slash	
sample standard deviation	100.0	(100.0)	11.4	(10.9)		
trimmed sd with $p = r = 0.2$	65.0	(89.9)	70.8	(100.0)	76.1	(28.1)
interquartile range	39.4	(41)	42.4	(47)	84.0	(94)
median absolute deviation	37.8	(35.3)	40.5	(41.5)	87.3	(91.8)
S_n	56.3	(54.1)				
Q_n	68.3	(68.8)				

Table 2.9. Comparison of one-pass efficiencies with those from published studies, shown in parentheses. Results for the trimmed standard deviation and median absolute deviation are compared with the results given in Lax (1985), the interquartile range with results given in Iglewicz (2000) for the fourth spread, and the results for S_n and Q_n to those given in Rousseeuw & Croux (1993). The latter efficiencies are based on standardised variances, rather than the variance of the log estimates. The one-pass efficiencies are averages based on 100 efficiencies, each from 20000 samples of size 20.

distribution, and very poor performance in the slash case. Theory tells us the first cannot be true since it is suboptimal to ignore the wild observation completely, and in fact by choosing $p = r = 0.2$ we ignore at least three well-behaved observations in each calculation. Also, the simulation results presented here show that elimination of a few extreme values in a slash sample results in a reasonable, but not perfect, level of efficiency. Results for the normal distribution cannot be affected by an understated numerator, so a likely conclusion is that Lax's trimmed standard deviations were not computed correctly.

The results for the MAD appear to correspond across all three distributions, and differences can almost certainly be attributed to sampling error and the vast differences in simulation sizes. Results for the IQR match the essence of those provided by Iglewicz for the fourth spread. Overstatement of the efficiencies for the IQR in the one-wild and slash cases could be due to an understated numerator in the efficiency calculation; however this is inconsistent with the MAD results (although these are from a different source).

Efficiencies for S_n and Q_n, based on standardised variances, are compared to those derived from the standardised variances given in Rousseeuw & Croux (1993) with only small differences observed. The original efficiencies were based on a single simulation run of 10000 samples of size 20. Thus, sampling error, and use of the rounded figures given by Rousseeuw & Croux (1993) could easily account for the differences.

The A-estimators

Lax found in favour of the A-estimators using the biweight ψ-function. However, alternative weight functions produced undominated estimators, in particular, the modified biweight and the modified sine functions. Since the Princeton Robustness Study (Andrews et al. 1972), the biweight has had periods of popularity in the robust literature (Cleveland 1979, Martinez & Iglewicz 1981, Kafadar 1982, Iglewicz & Martinez 1982, for example). It also features in a variety of currently popular robust techniques in particular the smoothing algorithm *loess* of Cleveland et al. (1992). For these reasons, we focus attention on A-estimators using the biweight ψ-function. In particular, optimal scaling constants for the modified biweight, and modified sine have not been sought. They are included here only for comparison with Lax's results.

Efficiencies of the A-estimators are compared in Figure 2.10, and average efficiencies are given in Table 2.10. It is immediately clear that the modified biweight lacks efficiency for the normal and one-wild distributions as was found in the Lax study. Performance of the modified sine is similar to that of the biweight with $c \in (9, 10, 11)$, in terms of triefficiency and also the range of efficiency across the three corners. The choices of scaling constant in the biweight estimators clearly demonstrate several phenomena. In theory, as $c \to \infty$, the biweight estimate converges to the sample standard deviation (albeit using the sample median rather than the sample mean), and hence we would expect to see its efficiency at the normal distribution increase with c. Also, as c increases, the point at which absolute deviations get zero weight increases, and hence we would expect to see the efficiency for both one-wild (for large c) and slash data decrease. Both these results are reflected in the simulated efficiencies presented here.

Unlike the one-pass estimators, the one-wild distribution dominates the triefficiencies for the A-estimators. In almost every case the efficiencies at the one-wild distribution are the lowest, and hence the triefficiency is simply the efficiency at the one-wild distribution. The intuition behind this is unclear; however, if we favour the use of these estimators, one interpretation is that the one-wild is not a suitable corner distribution when triefficiency is the criterion.

Despite the changes to both the numerator and denominator in the efficiencies, the biweight with $c = 9$ almost comes out as the triefficient estimator of this class. Use

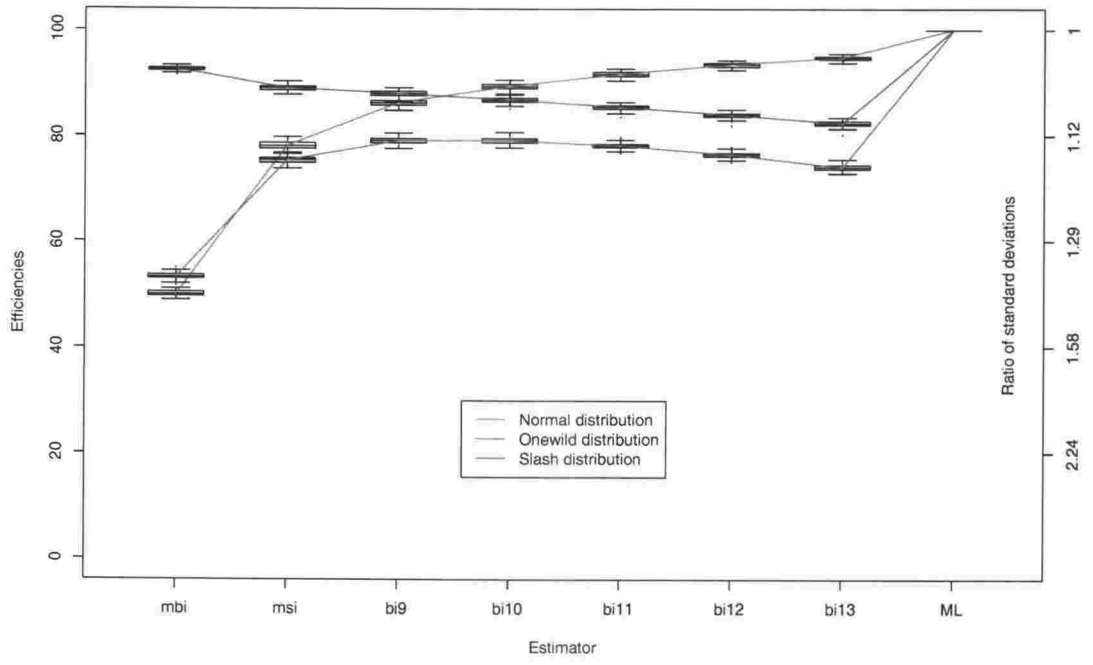

Figure 2.10. Efficiency distributions for the A-estimators using S_0 = MAD, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are mbi=modified biweight with $c = 6$, msi=modified sine with $c = 2.1$, bi9-bi13=biweight with constants $c = (9, 10, 11, 12, 13)$ respectively, and ML=maximum likelihood. Efficiency is computed using (2.31). The ratio of standard deviations is a non-linear scale given by (2.43).

of $c = 10$ improves the average one-wild efficiency and hence the average triefficiency by 0.1%. Should we consider only the normal and slash corners, the "biefficient" estimator would be the biweight with c between 9 and 10. The intersection of the lines joining the median efficiencies for the normal and slash distributions in Figure 2.10 suggests a scaling constant of close to 9.5 for maximum biefficiency across these two corners.

Once again there are conflicts between the results of this study and those of Lax. Average efficiencies are compared in Table 2.11 for the estimators common to both studies. In all cases, the figures for the normal data appear to correspond, and the general behaviour of efficiencies for both one-wild and slash data appear to correspond. However, whilst the order behaviour is generally consistent, point estimates cannot be recovered by rescaling Lax's efficiency estimates to account for overstated numerators, and hence again we conclude there are significant differences. In particular, the triefficiency of the best A-estimator is found to be smaller, at 79.2% rather than 85.8%.

estimator	normal	one-wild	slash	triefficiency
modified biweight with $c = 6$	50.0	53.3	92.5	50.0
modified sine with $c = 2.1$	78.1	75.3	89.0	75.3
biweight with $c = 9$	86.2	79.1	88.0	79.1
biweight with $c = 10$	89.4	79.2	86.8	79.2
biweight with $c = 11$	91.7	78.2	85.5	78.2
biweight with $c = 12$	93.4	76.5	84.0	76.5
biweight with $c = 13$	94.7	74.1	82.4	74.1

Table 2.10. Average efficiencies for the A-estimators using $S_0 = \text{MAD}$, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31). The triefficiency given is the average over the 100 simulations, rather than the minimum average. The efficiency distributions for these estimators are shown in Figure 2.10.

estimator	normal		one-wild		slash	
modified biweight with $c = 6$	50.0	(47.5)	53.3	(56.8)	92.5	(96.8)
modified sine with $c = 2.1$	78.1	(82.1)	75.3	(89.6)	89.0	(94.5)
biweight with $c = 9$	86.2	(86.7)	79.1	(85.8)	88.0	(86.1)
biweight with $c = 10$	89.4	(90.0)	79.2	(84.8)	86.8	(84.6)

Table 2.11. Comparison of A-estimator efficiencies with results from Lax (1985), shown in parentheses. The A-estimator efficiencies are averages based on 100 efficiencies, each from 20000 samples of size 20.

Estimators based on the t-distribution

An alternative to constructing an all-purpose estimator that achieves high trief-ficiency, such as an A-estimator, is to find an underlying compromise distribution whose corresponding (optimal) scale estimator has high triefficiency. Use of the Stu-dent's t distribution is one attempt to model this implicit distribution, which must be close to normal near the centre, exhibit intermediate tail behaviour, and have the possibility of the occasional "wild" observation. The t-estimators are multi-pass and as such need an initial scale estimate. If full iteration is performed, the choice of S_0 is not crucial and we can use the sample standard deviation. However, if only a single iteration is performed, we would hope to start with a scale estimate which is itself robust. Treating the one-step estimator as given in (2.29) as a special form of an A-estimator, we may use $S_0 = \text{MAD}$ and optimise by choice of scaling constant c.

Efficiencies of the fully iterated t-estimators are compared in Figure 2.11, and av-erage efficiencies are given in Table 2.12. Several surprising features emerge from these summaries. In particular, we note that the fully iterated t-estimator does par-ticularly well for the one-wild distribution when $\nu \in \{2, 3, 4\}$, but less well for the slash distribution. Even when $\nu = 1$ and the t-distribution is the Cauchy (compared to the slash in Table 2.1 and Figure 2.3) the optimal estimator for this distribu-tion only averages 76.8% efficiency for the slash. It is also interesting to note how quickly the normal efficiency increases with ν. In particular, even for $\nu = 6$, which would definitely be considered long-tailed and highly non-normal, the t-estimator is nearly 95% efficient for the normal samples. The one-wild performance when $\nu = 3$ is the second best of all estimators considered in the simulation, following the trimmed standard deviation with $p = r = 0.1$ which has an average efficiency of 88.1%. Normal/one-wild biefficiency is achieved between $\nu = 2$ and 3, which seems surprisingly low.

Efficiencies of the one-step t-estimators are compared in Figure 2.12, and average efficiencies are given in Table 2.13. We see that like the biweight A-estimator and the one-pass estimators, this estimator has failed to cope well with the one-wild samples. As c, and hence the implicit degrees of freedom ν, decreases, the t-distribution behaves much like the slash distribution in the tails, and as expected, the efficiency increases. Similarly as ν increases, efficiency increases for the normal data, which is of course the distribution obtained as $\nu \to \infty$. The best choice of c appears to

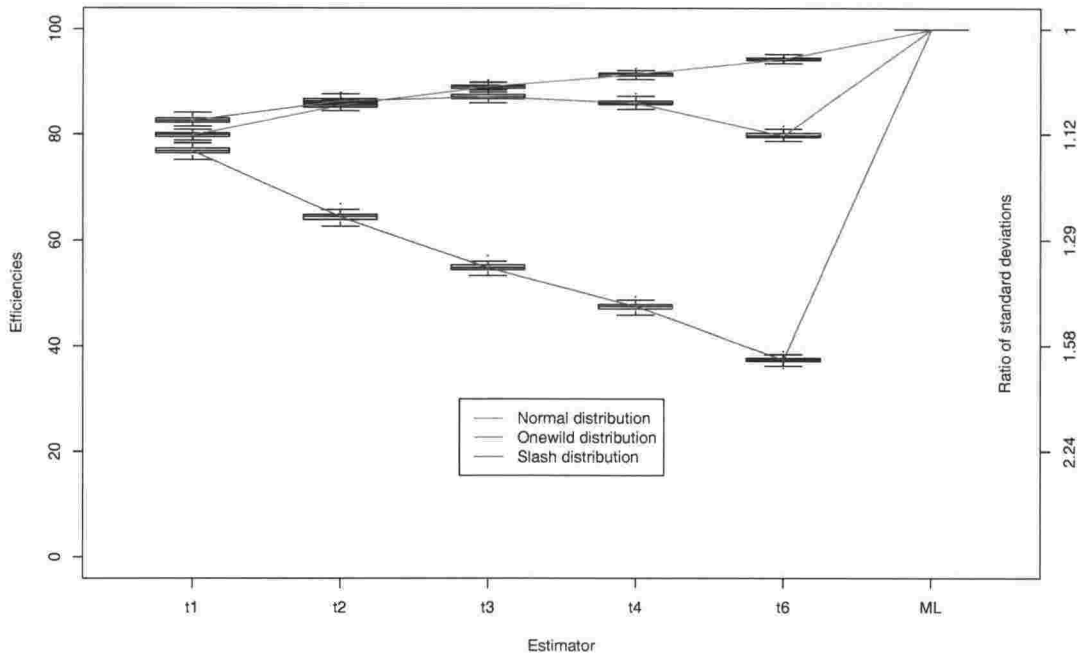

Figure 2.11. Efficiency distributions for the fully iterated t-estimators, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators t1-t6 have associated degrees of freedom $\nu = (1, 2, 3, 4, 6)$ respectively, and ML=maximum likelihood. Efficiency is computed using (2.31). The ratio of standard deviations is a non-linear scale given by (2.43).

estimator	normal	one-wild	slash	triefficiency
fully iterated t with $\nu = 1$	79.8	82.6	76.8	76.8
fully iterated t with $\nu = 2$	85.5	86.3	64.3	64.3
fully iterated t with $\nu = 3$	89.0	87.1	54.9	54.9
fully iterated t with $\nu = 4$	91.4	86.0	47.4	47.4
fully iterated t with $\nu = 6$	94.4	79.8	37.4	37.4

Table 2.12. Average efficiencies for the fully iterated t-estimators, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31). The triefficiency given is the average over the 100 simulations, rather than the minimum average. The efficiency distributions for these estimators are shown in Figure 2.11.

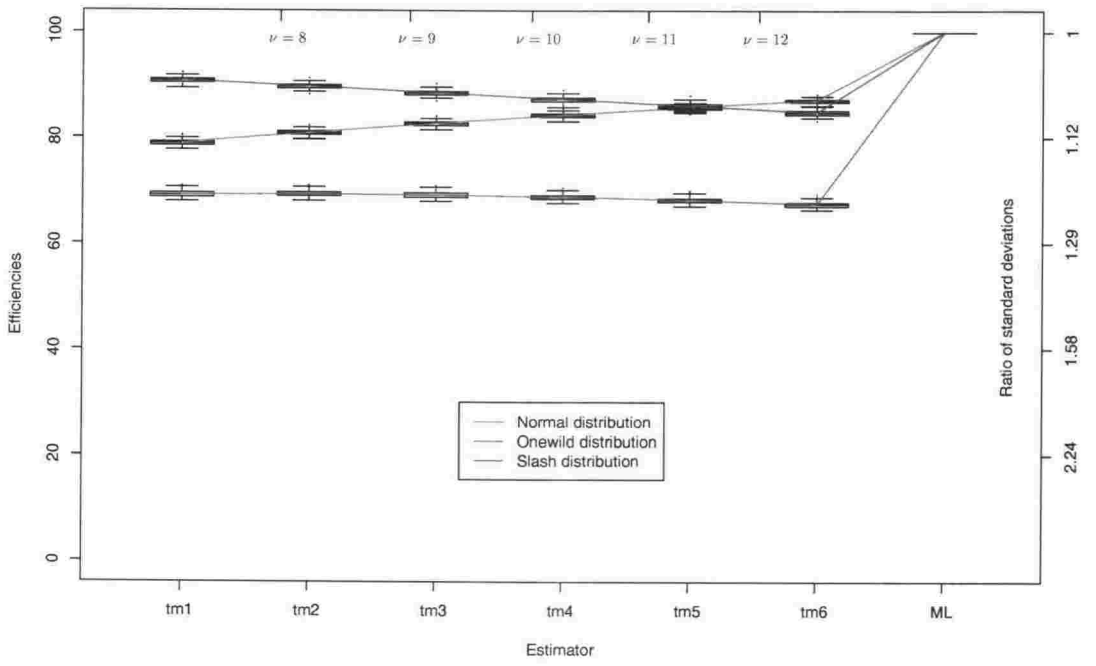

Figure 2.12. Efficiency distributions for the one-step t-estimators, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators tm1-tm6 use S_0 = MAD and have associated scaling parameters $c = (4, 4.25, 4.5, 4.75, 5, 5.25)$ respectively, and ML=maximum likelihood. The (non-linear) upper axis shows the corresponding size of ν if $S_0 = 1.4826$(MAD) is assumed. Efficiency is computed using (2.31). The ratio of standard deviations is a non-linear scale given by (2.43).

be $c = 4.25$ using a triefficiency criterion, or $c = 5$ if biefficiency for the normal and slash distributions is sought. These values of c correspond roughly to $\nu = 8$ and $\nu = 12$ respectively. It is interesting that lower degrees of freedom are needed when normal/one-wild biefficiency is sought, even though the tail of the one-wild distribution is better behaved than that of the slash. Also, we note that using the MAD and $c = 4$ we obtain an average slash efficiency higher than any of the fully iterated t-estimators and 3.5% higher than the MAD itself. Overall, the t-estimators do not appear to be as good as the biweight A-estimators: their performance at the one-wild is approximately 10% worse, and their normal/slash biefficiency is marginally lower. Further, use of the MAD and a single iteration fails to provide the triefficiency of the fully iterated estimator with $\nu = 1$.

These results are intriguing in the sense that we see quite different behaviour from the fully iterated and one-step t-estimators. In particular, the fully iterated estimators do well for the normal and one-wild corners but relatively poorly for the slash data, despite low values of ν. In contrast, for much higher implicit values of ν, the one-step estimators do very well for slash data, and less well for the normal and one-wild samples. The t-distribution does appear to be able to moderate between

estimator	normal	one-wild	slash	triefficiency
one-step t with $c = 4.00$	78.9	69.1	90.8	69.1
one-step t with $c = 4.25$	80.8	69.3	89.7	69.3
one-step t with $c = 4.50$	82.6	69.1	88.5	69.1
one-step t with $c = 4.75$	84.3	68.8	87.3	68.8
one-step t with $c = 5.00$	85.7	68.2	86.1	68.2
one-step t with $c = 5.25$	87.0	67.4	84.8	67.4

Table 2.13. Average efficiencies for the one-step t-estimators using S_0 = MAD, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31). The triefficiency given is the average over the 100 simulations, rather than the minimum average. The efficiency distributions for these estimators are shown in Figure 2.12.

the corners, and perhaps introduction of truncating weights for the fully iterated case may improve the slash performance, without too much loss in efficiency at the remaining corners. We leave this a subject for future research.

2.5.3 Use of alternative auxiliary estimates

The results of the Section 2.5.2 clearly show that the estimators of Rousseeuw & Croux (1993) (i.e. S_n and Q_n), are more efficient at each of the three corners than the median absolute deviation, and indeed Q_n has the highest average triefficiency of any of the single-pass estimators considered in this study. This suggests that use of either S_n or Q_n as auxiliary scale estimators in multiple-pass estimators, could increase the efficiency of those classes of estimators.

In order to use either S_n or Q_n in the biweight A-estimator, the scaling constant c must also be changed. We consider using S_n with $c = (6.5, 7, 7.5)$ and Q_n with $c = (10.5, 11, 11.5)$. These choices seem to provide a maximum for the one-wild efficiency as shown in Figure 2.13 and Table 2.14. Evident is the gain in efficiency of the A-estimator from using the more efficient auxiliary estimators of scale. In particular, use of Q_n and a scaling constant of 11 increases the one-wild efficiency

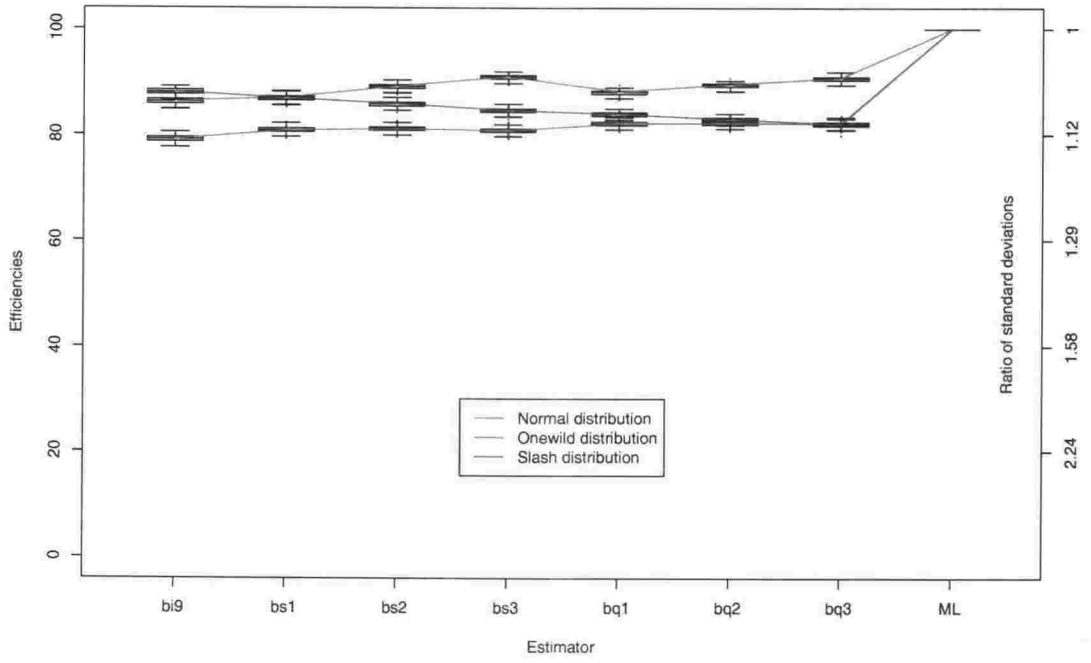

Figure 2.13. Efficiency distributions for biweight A-estimators with alternative S_0, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are bi9=biweight with MAD and $c = 9$, bs1-bs3 use S_n with $c = (6.5, 7, 7.5)$ respectively, bq1-bq3 use Q_n with $c = (10.5, 11, 11.5)$ respectively, and ML=maximum likelihood. Efficiency is computed using (2.31). The ratio of standard deviations is a non-linear scale given by (2.43).

estimator	normal	one-wild	slash	triefficiency
biweight with MAD and $c = 10$	89.4	79.2	86.8	79.2
biweight with S_n and $c = 6.5$	86.8	80.8	86.9	80.8
biweight with S_n and $c = 7$	89.0	81.1	85.8	81.1
biweight with S_n and $c = 7.5$	90.8	80.8	84.6	80.8
biweight with Q_n and $c = 10.5$	88.0	82.1	83.9	82.1
biweight with Q_n and $c = 11$	89.4	82.2	82.9	82.1
biweight with Q_n and $c = 11.5$	90.6	82.1	82.0	81.7

Table 2.14. Average efficiencies for the A-estimators using alternative S_0, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31). The triefficiency given is the average over the 100 simulations, rather than the minimum average. The efficiency distributions for these estimators are shown in Figure 2.13.

	normal	one-wild	slash
MAD	0.6473	0.6852	1.5067
S_n	0.8582	0.9266	2.1878
Q_n	0.5360	0.5895	1.4913
10 MAD	6.4727	6.8524	15.0672
$7S_n$	6.0073	6.4860	15.3147
$11Q_n$	5.8957	6.4850	16.4045

Table 2.15. Simulation average of the auxiliary scale estimates and the corresponding averages of those estimates times the best selected scaling constants.

choices of cS_0 in the A-estimators. From this table it is clear why the efficiencies have decreased for the slash data, since by increasing cS_0 we would expect these empirical responses as fewer data are eliminated by the biweight function. The scaling constants have essentially been determined by the one-wild performance, and this improvement is reflected in the decrease of cS_0. An interesting effect is that the normal efficiency does not decline as a result of cS_0 decreasing for that corner as well.

Having noted that use of more efficient auxiliary scale estimators has benefitted the A-estimators, the same principles also hold true for the one-step estimators based on the t-distribution. Since they are not fully iterated, choice of a "better" initial scale estimate may improve the efficiency of the final estimate. Use of ν as the scaling constant would require a "good" initial estimate. Focussing on consistency in the normal case, and taking scaling constants from Table 2.15, we would typically use $S_0 = \text{MAD}/0.6473$, $S_0 = S_n/0.8582$ or $S_0 = Q_n/0.5360$ depending on our preference for MAD, S_n or Q_n. Since we have reparameterised the one-step t-estimators, this is unnecessary and we simply choose MAD, S_n or Q_n and maximise the triefficiency by choice of c. We simulate using S_n with $c \in \{2.75, 3, 3.25\}$ and Q_n with $c \in \{4, 4.25, 4.5\}$. The results of these simulations are summarised in Table 2.16 and Figure 2.14, along with results for the best fully iterated t-estimator and one-step using the MAD. The results clearly show an improvement in overall efficiency.

Use of S_n results in triefficiency comparable to the fully-iterated t, with lower performance at the one-wild, but higher efficiency at both the normal and slash distributions. Further improvement is achieved by using Q_n, with a small compromise in normal-slash efficiencies for an improvement in one-wild efficiency. Once again the influence of the one-wild results is seen. Maximum efficiency for the one-wild occurs

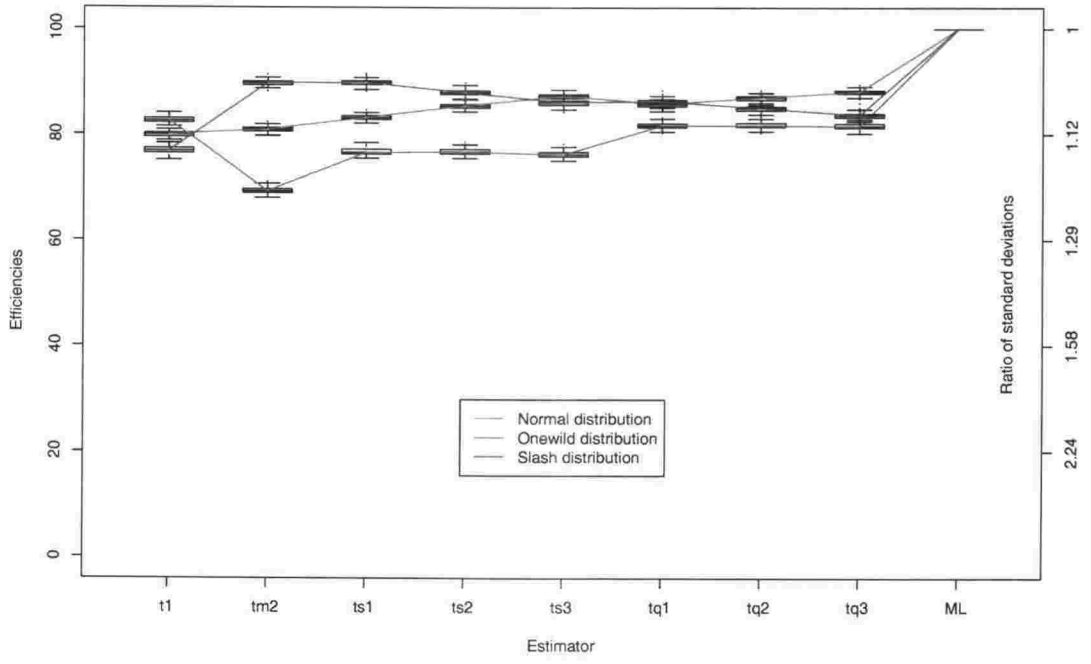

Figure 2.14. Efficiency distributions for t-estimators with alternative S_0, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are t1=fully iterated with $\nu = 1$, tm2=one-step with $S_0 = \text{MAD}$ and $c = 4.25$, ts1-ts3 are one-step with $S_0 = S_n$ and $c = (2.75, 3, 3.25)$, tq1-tq3 are one-step with $S_0 = Q_n$ and $c = (4, 4.25, 4.5)$ respectively, and ML=maximum likelihood. Efficiency is computed using (2.31). The ratio of standard deviations is a non-linear scale given by (2.43).

at lower c when Q_n is used, but despite this, efficiencies for the normal *increase* and those for the slash *decrease*. We would expect the opposite effect simply by decreasing the degrees of freedom, but there is an offsetting effect due to the relative sizes of S_0 using the MAD, S_n and Q_n similar to the effect illustrated in Table 2.15 for the biweight A-estimators.

The t-estimator results are now very similar across all three distributions to those for the A-estimator with Q_n and $c = 11$. We see slightly worse performance for the one-step t with Q_n and $c = 4.25$ at the normal and one-wild, better performance at the slash, and a 0.3% decrease in average triefficiency. Thus the one-step t-estimator with Q_n and $c = 4.25$ emerges as a serious contender for the biweight A-estimator with Q_n and $c = 11$.

2.5.4 Other results

Alternative sample sizes

We repeat some of the above analysis for samples of sizes 10 and 40. We would expect the results for the one-wild to exhibit the greatest changes here, and dominate the

estimator	normal	one-wild	slash	triefficiency
fully iterated t with $\nu = 1$	79.8	82.6	76.8	76.8
one-step t with MAD and $c = 4.25$	80.8	69.3	89.7	69.3
one-step t with S_n and $c = 2.75$	83.1	76.6	89.8	76.6
one-step t with S_n and $c = 3$	85.3	76.6	87.9	76.6
one-step t with S_n and $c = 3.25$	87.3	76.2	86.0	76.2
one-step t with Q_n and $c = 4$	85.7	81.7	86.2	81.7
one-step t with Q_n and $c = 4.25$	86.9	81.8	85.0	81.8
one-step t with Q_n and $c = 4.5$	88.1	81.7	83.7	81.7

Table 2.16. Average efficiencies for the t-estimators using alternative S_0, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31). The triefficiency given is the average over the 100 simulations, rather than the minimum average. The efficiency distributions for these estimators are shown in Figure 2.14.

results for the smaller samples particularly. In the case of $n = 10$, the single "wild" observation represents a greater proportion of the sample, and we would expect efficiencies at the one-wild to decrease for most estimators. However, when $n = 40$, the wild observation should not be so dominant, and we would expect estimators to have efficiencies much closer to their Gaussian efficiencies. These effects will have a direct impact on triefficiencies due to the dominance of the one-wild distribution in the results for $n = 20$.

Results are presented for selected estimators in Tables 2.17 to 2.19 for the normal, one-wild and slash distributions respectively. Consider first the results given for the normal distribution in Table 2.17. Here, no systematic behaviour is observed across all estimators. The average efficiencies of the trimmed standard deviation and MAD decrease as sample size increases, whereas S_n, Q_n and the biweight with both MAD and S_n perform better relative to the sample standard deviation. Interestingly, we see different behaviour for MAD (whose preformance worsens as n increases) and its proposed alternatives S_n and Q_n (whose performance increases). Further, the average efficiencies of the biweight with Q_n and scaling constants $c = 10.5$ and $c = 11$ are approximately constant whereas the one-step t with Q_n and $c = 4.25$ does slightly worse as sample size increases. These latter three estimators do very well at all sample sizes with average efficiencies close to 90%.

The results for the one-wild distribution are shown in Table 2.18. Intuitively we would expect the efficiencies to increase with sample size for the robust estimators, and this is exactly what is observed. The sample standard deviation gets worse as sample size increases, as we would expect for an estimator with no protection

estimator	$n = 10$	$n = 20$	$n = 40$
sample standard deviation	100.0	100.0	100.0
trimmed sd with $p = r = 0.1$	86.5	80.9	79.7
median absolute deviation	39.1	37.8	37.4
S_n	50.2	54.7	57.9
Q_n	60.7	66.9	72.7
biweight with MAD and $c = 9$	72.7	86.2	90.3
biweight with MAD and $c = 10$	77.6	89.4	92.6
biweight with S_n and $c = 6.5$	78.5	86.8	89.0
biweight with S_n and $c = 7$	81.6	89.0	90.9
biweight with Q_n and $c = 10.5$	87.8	88.0	87.0
biweight with Q_n and $c = 11$	89.2	89.4	88.5
one-step t with Q_n and $c = 4.25$	88.2	86.9	86.5

Table 2.17. Average efficiencies for selected estimators for the normal distribution only, based on 100 realisations of the efficiencies, each estimated from 20000 samples of sizes 10, 20 and 40. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31).

against the single "wild" observation. The trimmed standard deviation has an aberrant value for $n = 10$ because the two-sided trimmed mean is computed for the entire sample (the trimming in this case removes 5% of the observations from each end of the ranked sample, and here $0.05n < 1$ so no observations are trimmed). Relative performance worsens for $n = 40$ because of the parameterisation of the estimator. Since we remove $100r\%$ of the observations, as n increases the number of "good" observations trimmed increases, and consequently efficiency decreases. MAD gets worse as n increases; however both S_n and Q_n improve. The A-estimators and the one-step t-estimator all show a systematic increase in efficiency as sample size increases, as we would expect. Here, the identification and down-weighting of the single "wild" observation improves as n increases, and the scale estimates are computed with similar efficiency to those for the normal distribution and $n = 40$. While these estimators were only able to attain one-wild efficiencies close to 80% for $n = 20$, when $n = 40$, the efficiencies are higher, and much closer to those attained for the normal distribution.

The results for the slash distribution are shown in Table 2.19. Once again, as we would expect, the sample standard deviation gets worse as sample size increases. Unlike its performance for one-wild samples, the trimmed standard deviation benefits in the slash case from losing a larger number of observations, and increases in efficiency as n increases (once again the figure for $n = 10$ is deflated). The MAD and S_n become less efficient as n increases, whereas the performance of Q_n is fairly

estimator	$n = 10$	$n = 20$	$n = 40$
sample standard deviation	17.8	11.4	8.6
trimmed sd with $p = r = 0.1$	50.6	88.1	83.9
median absolute deviation	44.4	40.5	39.0
S_n	53.2	55.9	58.7
Q_n	63.6	68.4	73.7
biweight with MAD and $c = 9$	63.5	79.1	86.9
biweight with MAD and $c = 10$	63.6	79.2	87.5
biweight with S_n and $c = 6.5$	66.4	80.8	86.8
biweight with S_n and $c = 7$	65.8	81.1	87.7
biweight with Q_n and $c = 10.5$	66.1	82.1	85.9
biweight with Q_n and $c = 11$	64.7	82.2	86.8
one-step t with Q_n and $c = 4.25$	72.7	81.8	85.4
maximum likelihood	100.0	100.0	100.0

Table 2.18. Average efficiencies for selected estimators for the one-wild corner only, based on 100 realisations of the efficiencies, each estimated from 20000 samples of sizes 10, 20 and 40. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31).

constant. The A-estimators, and the one-step t, all do worse as n increases and comparison with Tables 2.17 and 2.18 shows that the triefficiencies for these estimators at $n = 40$ are now their slash efficiencies.

For $n = 10$, the one-step t with Q_n and $c = 4.25$ is the triefficient estimator, due to its 72.7% average efficiency at the one-wild distribution (as shown in Figure 2.15). We also note that this is the biefficient estimator for the normal and slash corners for this particular sample size. No attempt was made to optimise c for any of the two-pass estimators, so there may be better choices when $n = 10$. For $n = 40$, the efficiencies are shown in Figure 2.16. In this plot it is clear that, for $n = 40$, the trimmed standard deviation, MAD, S_n and Q_n are not greatly influenced by the "wild" observation in the one-wild samples. The best estimator is the biweight with MAD and $c = 9$, with an average triefficiency of 86.2%. Like the one-step t when $n = 10$, this was the best performing estimator in all 100 trials. The biweight with MAD and $c = 10$, with S_n and $c = 6.5$, and the one-step t also had high triefficiencies, with averages 84.8%, 84.7% and 84.6% respectively. In fact, there is very little difference between the two-step estimators considered.

Results using standardised variances

Use of the standardised variance favoured by Rousseeuw & Croux (1993), giving the efficiency (2.32), rather than log-variance used by Lax, with efficiency given by

estimator	$n = 10$	$n = 20$	$n = 40$
sample standard deviation	17.4	7.5	3.5
trimmed sd with $p = r = 0.1$	26.2	42.1	43.3
median absolute deviation	92.3	87.3	85.4
S_n	98.3	95.8	94.5
Q_n	96.8	94.9	95.1
biweight with MAD and $c = 9$	90.7	88.0	86.3
biweight with MAD and $c = 10$	90.4	86.8	84.8
biweight with S_n and $c = 6.5$	90.7	86.9	84.7
biweight with S_n and $c = 7$	90.0	85.8	83.4
biweight with Q_n and $c = 10.5$	86.2	83.9	83.2
biweight with Q_n and $c = 11$	85.4	82.9	82.3
one-step t with Q_n and $c = 4.25$	87.3	85.0	84.7
maximum likelihood	100.0	100.0	100.0

Table 2.19. Average efficiencies for selected estimators for the slash distribution only, based on 100 realisations of the efficiencies, each estimated from 20000 samples of sizes 10, 20 and 40. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31).

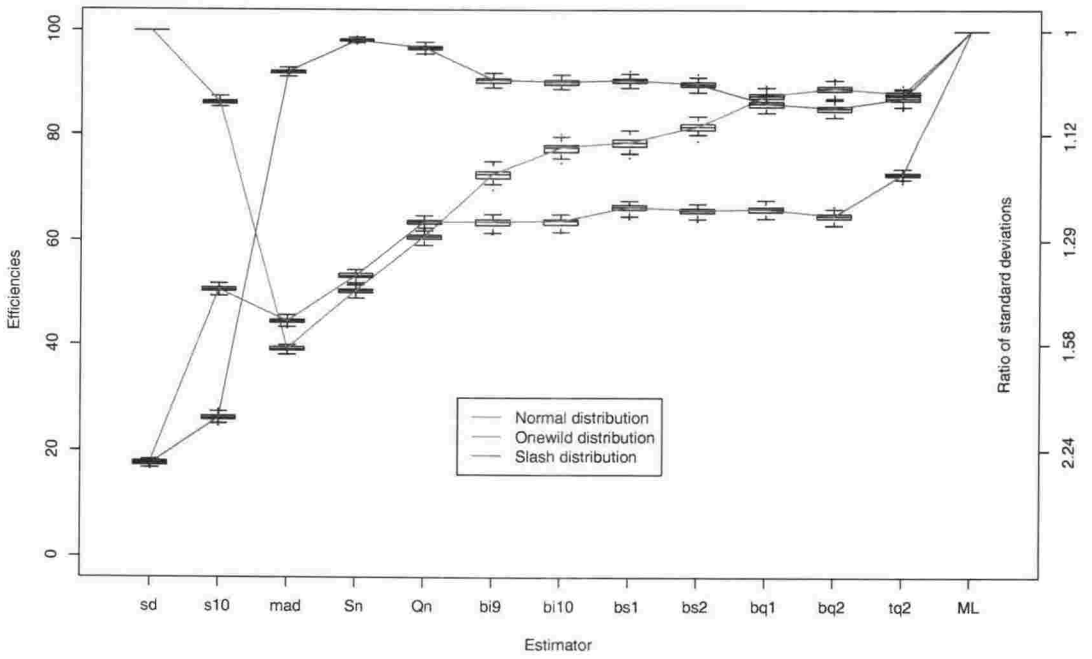

Figure 2.15. Efficiency distributions for estimators with $n = 10$, based on 100 realisations of the efficiencies, each estimated from 20000 samples. The estimators are sd=sample standard deviation, s10=trimmed sd with $p = r = 0.1$, mad=MAD, Sn=S_n, Qn=Q_n, bi9-bq2=biweight with MAD and $c = (9, 10)$, S_n and $c = (6.5, 7)$ and Q_n with $c = (10.5, 11)$ respectively, tq2=one-step t with Q_n and $c = 4.25$ and ML=maximum likelihood. Efficiency is computed using (2.31). The ratio of standard deviations is a non-linear scale given by (2.43).

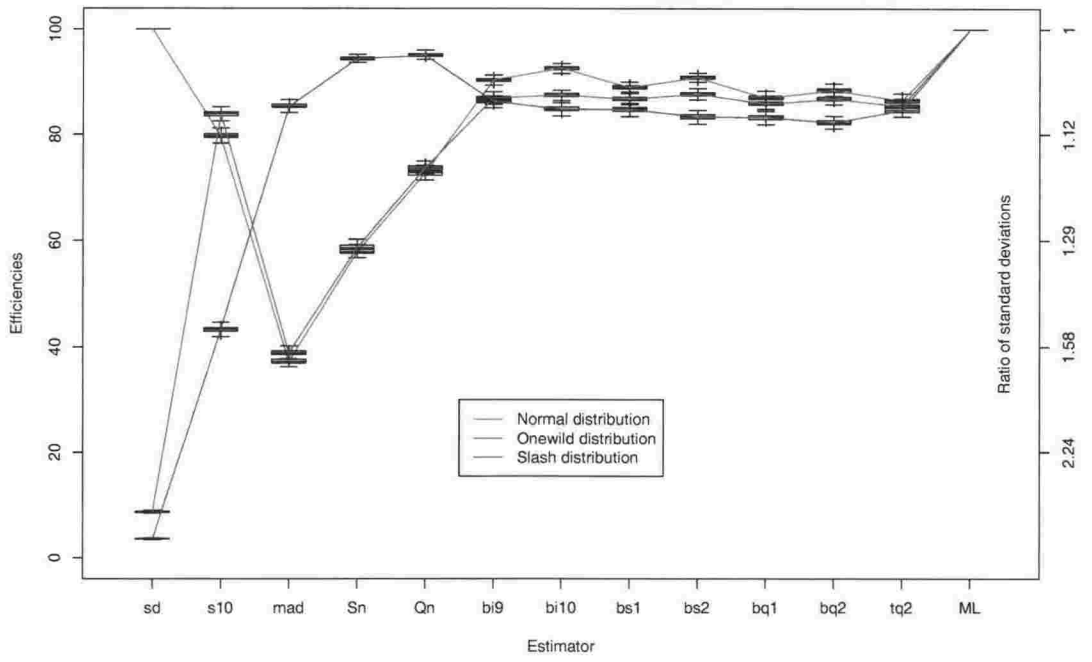

Figure 2.16. Efficiency distributions for estimators with $n = 40$, based on 100 realisations of the efficiencies, each estimated from 20000 samples. The estimators are sd=sample standard deviation, s10=trimmed sd with $p = r = 0.1$, mad=MAD, Sn=S_n, Qn=Q_n, bi9-bq2=biweight with MAD and $c = (9, 10)$, S_n and $c = (6.5, 7)$ and Q_n with $c = (10.5, 11)$ respectively, tq2=one-step t with Q_n and $c = 4.25$ and ML=maximum likelihood. Efficiency is computed using (2.31). The ratio of standard deviations is a non-linear scale given by (2.43).

(2.31), has an interesting effect on the results for $n = 20$. The average efficiencies based on standardised variances are found in Table C.2 and these can be compared to the complete results based on log variances given in Table C.1. The estimators in both tables are sorted according to average rank using the triefficiency as the criterion. In almost every case, average efficiency at the normal distribution is higher using the standardised variances than it was using the variance of the log estimates. The exceptions are Gini's mean difference, and the fully iterated t-estimators with $\nu = 4$ and 6, each with an average difference very close to zero. The largest gain is 4.1% for the modified biweight A-estimator, but most estimators gain less than 1% efficiency.

The one-wild results lose some influence on the triefficiencies under the alternative efficiency estimate, and do not exhibit large systematic changes like the normal and slash results. Nearly all the efficiencies for the single-pass estimators increase, whereas the multi-pass estimators generally lose ground. Most average changes for this class are very small, with only the biweights using MAD and $c = 11$, 12, and 13 falling more than 1%.

The efficiencies for the slash distribution show an opposite effect to the normal results; here most estimators have lower efficiencies using the standardised variances, and unlike the normal data, the differences in some cases are sizeable. This is clearly related to the stabilising effect of the log transform. The largest differences are observed in the estimators that perform poorly at the slash distribution: sample standard deviation, Gini's mean difference, the trimmed standard deviations, and the fully iterated t-estimators. All of these differences (except the standard deviation's) are above 10%, with the trimmed standard deviation with $p = r = 0.1$ falling almost 25%. All other estimators maintain average slash efficiencies in excess of 75%, as seen in Table C.2. In two instances (the MAD and the modified biweight) the slash efficiencies increase by approximately 1%, however most changes are decreases of between 3 and 6%. Substantial drops for the two-pass estimators using Q_n as an auxiliary scale (of around 6%) cause the slash efficiency to be the lowest of the three distributions for these estimators. This affects the biweights using S_n and Q_n, and the one-step t using Q_n in particular. Interestingly, these estimators are among the best performing estimators under the measure based on log variance. Under that measure, the best triefficiencies are slightly higher, at around 82%.

Since the one-wild results are fairly stable, so too are the triefficiencies. We do however select different estimators on account of large decreases in slash efficiency for some estimators. The best performing estimators are now the biweight with S_n and $c = 6.5$, $c = 7$, and the one-step t with Q_n and $c = 4$. All three estimators have triefficiencies in excess of 80%, as seen in Table C.2. This effect is shown graphically in Figure 2.17, where the code for each estimator is plotted against average rank using log variance (on the horizontal axis) and average rank based on standardised variance (on the vertical axis). The cluster of estimators in the lower left quadrant of the plot indicate those which have performed well under both measures, and the minimum average triefficiency of this group is 75.9% for the biweight with Q_n and $c = 11.5$ (bq3). In short, use of the standardised variance does not have a very dramatic effect on the conclusions we draw.

2.6 Conclusions

A large simulation has been performed, and has produced results in conflict with those of Lax (1985). In particular, the estimator with the largest triefficiency iden-

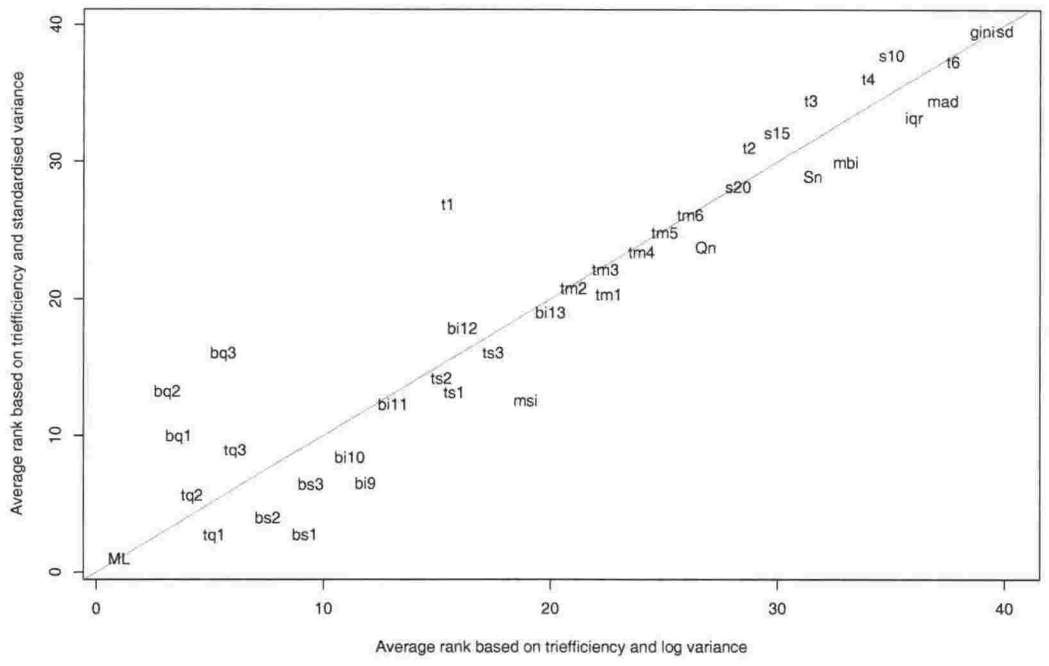

Figure 2.17. Comparison of average ranks under the two efficiency measures. The average ranks based on triefficiency and the variance of the log estimates is on the horizontal axis, and the average ranks based on triefficiency and the standardised variance (2.32) on the vertical axis, and the estimator is indicated by its short code (see Table 2.6). The line representing equality is shown in grey, and full results are given in Tables C.1 and C.2.

tified by Lax is not the estimator with greatest triefficiency in this study, and the triefficiency found by Lax of 85.8% has not been reached.

Figure 2.18 compares the efficiencies of the best estimators of each class: the biweight with MAD and $c = 10$, the biweight with S_n and $c = 7$, the biweight with Q_n and $c = 11$, and the t-estimator with Q_n and $c = 4.25$. From this plot it is clear that these four estimators offer estimates of very similar quality. Even though the gains in efficiency are likely to be small, use of the biweight with Rousseeuw & Croux's (1993) S_n or Q_n, or use of the t-estimator with Q_n and a single iteration, will provide better estimates than use of the biweight with MAD and $c = 10$ (and therefore also with MAD and $c = 9$). Except for the biweight with Q_n, all triefficiencies for these four estimators are based exclusively on the one-wild distribution, and other estimators may be preferred if the triefficiency ceases to be the selection criterion, i.e., if a different distribution to the one-wild is chosen as the third corner.

All efficiency distributions have similar ranges and interquartile ranges and are reasonably symmetric. The efficiencies for any particular estimator and distribution combination typically have standard deviations close to 0.5%. Thus the standard

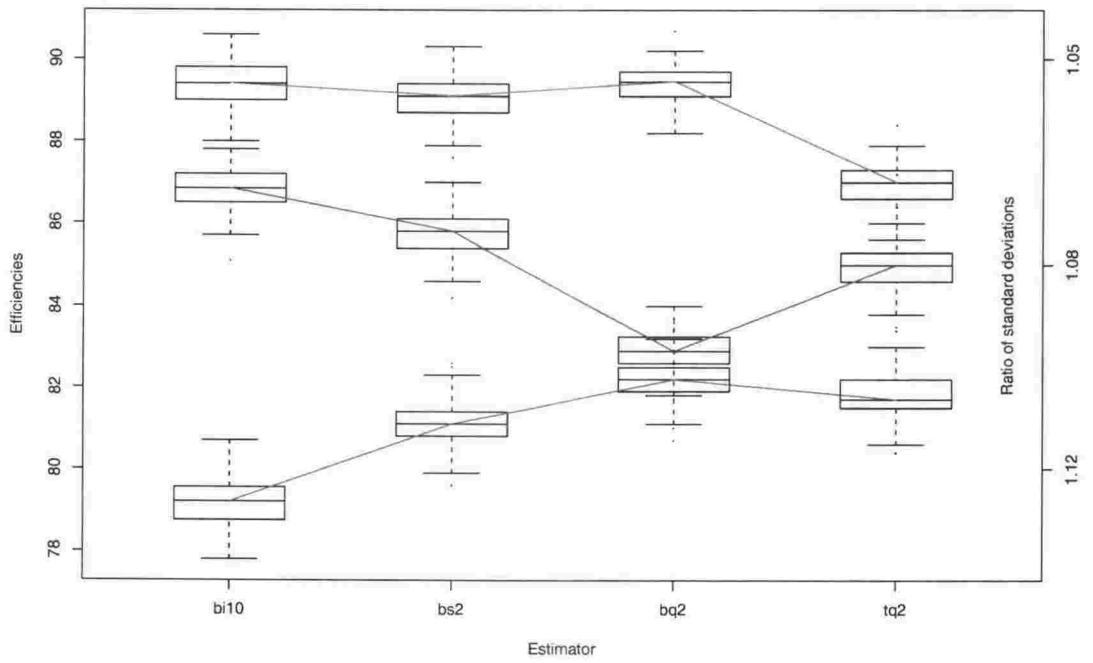

Figure 2.18. Efficiency distributions for the best performing estimators of each class, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are bi10=biweight A-estimator with MAD and $c = 10$, bs2=biweight with S_n and $c = 7$, bq2=biweight with Q_n and $c = 11$, and tq2=the t-estimator with Q_n and $c = 4.25$. The red lines join the medians for the normal distribution, the green the medians for the one-wild, and the blue for the slash. Efficiency is computed using (2.31). The ratio of standard deviations is a non-linear scale given by (2.43).

errors of the average efficiencies are approximately 0.05%, and we can assume that the reported average efficiencies are very close indeed to the true efficiencies.

Even though the very long tails of the slash distribution would seem to present a greater challenge, relatively speaking, high efficiency at the one-wild has proven a more difficult achievement for an estimator which performs well in general. The sampling variance of the estimates in the slash situation is of course much higher than that in the one-wild case. However, robust estimators struggle to match the performance of the maximum likelihood estimates for the one-wild in particular. The maximum average efficiency for any estimator considered was 98.0% for Gini's mean difference at the normal distribution, 95.8% for S_n at the slash distribution, but only 88.1% for the trimmed standard deviation with $p = r = 0.1$ at the one-wild. This does raise the question as to whether or not the three corners used for the triefficiency are indeed the appropriate corners. A simple modification would be to lower the standard deviation of the wild observation to perhaps eight or nine times that of the others, and this would raise the efficiency of many estimators for

the one-wild samples. An alternative would be to stay with $k = 10$, but treat this as an unknown parameter in the maximum likelihood recursions.

By undertaking a study of this magnitude, it is intended that the efficiencies reported will become the benchmark for robust scale estimators. However, many avenues exist to extend these results. In particular, no attempt was made to examine alternative ψ-functions to the biweight in the A-estimators. Further, the parameter k in Q_n was treated as fixed. Adjustment of this might lead to greater efficiency both for Q_n and for estimators using it as an auxiliary scale estimator. The t-estimators performed well despite having non-zero weights for all observations, and with Q_n and $c = 4.25$ was both the normal/slash biefficient estimator and the triefficient estimator when $n = 10$. Truncation of the weight function of the t-estimator at some suitable point might improve the performance of these estimators.

Together with the results for three location estimators presented in Appendix B, it is hoped that this work will prompt further interest in robust scale (and location) estimation generally, in particular, regarding the performance for small samples.

Chapter 3

Non-parametric volatility estimation

Evolving volatility is a dominant feature observed in most financial time series and a key parameter used in option pricing and many other financial risk analyses. Although there is now an extensive literature on the estimation of parametric volatility models (see Engle (1982), Taylor (1986), Bollerslev, Chou & Kroner (1992), Harvey, Ruiz & Shephard (1994), Bollerslev & Mikkelson (1996), Shephard (1996) and Barndorff-Nielsen & Shephard (2001) for example) less attention has been paid to simpler non-parametric alternatives. Exceptions include Aït-Sahalia (1996), Anderson & Grier (1992), and Andersen, Bollerslev, Diebold & Labys (2001) for example. More closely related to this paper is the work of Turner & Weigel (1992) who analyse the volatility of the daily returns of the S&P 500 and Dow Jones indices using the sample interquartile range (see Definition 2.8) as well as other measures of volatility. However, such estimates need to be rescaled in order to provide an unbiased estimate of the standard deviation of the underlying data since this is the predominant measure of volatility in financial applications, due to its use in standard option pricing and portfolio optimisation methodologies.

In this chapter we present preliminary findings on the construction and properties of non-parametric estimators of time-varying volatility, where volatility is assumed to be a measure of scale. Our focus is on financial data and, in particular, the daily returns of market prices such as equities, market indices and exchange rates, where daily returns are the first differences of the logarithms of the prices. Thus each daily return measures the continuously compounded rate of return on the asset or index over the day concerned. It is noted that non-parametric volatility estimators are particularly appropriate for extracting and understanding historical volatility

prior to fitting a more sophisticated parametric model. They also provide robust benchmarks for testing the forecasting and in-sample performance of competing parametric procedures.

Our objective is to construct non-parametric volatility estimators that have simple structure, are cheap to compute, and are tailored to the typically heavy-tailed distributions met in practice. In particular we seek procedures that are robust to distributional assumptions, resistant to outliers, and have a sound statistical basis with reasonable precision properties. The estimation procedures that we consider construct local robust scale estimates (not necessarily estimates of standard deviation) based on finite moving-averages of the squared deviations of the time series from its local level. The moving-average weights are selected with reference to a target family of heavy-tailed distributions, and the span of the moving-average is chosen so that the volatility is approximately constant within the local time window concerned. Finally, a global correction factor is applied to the local scale estimates to provide estimates of time-varying volatility.

No attempt has been made to build a predictive model, and, as with other window-based estimators, there are issues to be addressed as to what to do at the ends of the series.

We choose to model a time series of (daily) returns R_t as

$$R_t = \mu_t + \sigma_t \epsilon_t \tag{3.1}$$

where $R_t = \ln S_t - \ln S_{t-1}$, time t is measured in days, and S_t is the underlying time series of prices concerned. The ϵ_t are assumed to be independently and identically distributed with mean zero and unit variance. The condition $\mathrm{E}(\epsilon_t^2) = 1$ serves to identify the volatility σ_t which is assumed to be a strictly positive, smoothly-varying function of time, so that R_t has mean μ_t and variance σ_t^2. For daily data, μ_t will typically be very small in relation to $\sigma_t \epsilon_t$, since it represents the mean return over just one day. However, we shall assume that R_t has been corrected for such an evolving mean level if appropriate. The R_t will typically be heavy-tailed so that the common distribution of the ϵ_t will be heavy-tailed also.

3.1 Discussion of assumptions

The non-parametric estimator of volatility we propose is intended to provide a modern method of estimating historical volatility based on a suitably chosen finite mov-

ing average of the squared mean-corrected daily returns. It is based on a local model, rather than a global parametric model, and is designed to give robust and resistant estimates with good efficiency properties that could be used to aid in model selection, and to benchmark forecasts for global parametric volatility models. In order to facilitate our estimator we need to make some minimal basic assumptions, and these are discussed below.

The first key assumption is that the mean μ_t and volatility σ_t generally change smoothly over time, and in particular, are locally constant over the local time windows within which estimation takes place. This seems a reasonable assumption for the most part and without it reliable volatility estimation would be difficult to achieve.

The smoothness of volatility is embodied in the slow decay or long memory of the autocorrelation function of absolute stock price returns, as demonstrated by Ding, Granger & Engle (1993), Granger & Ding (1995), Rydén, Teräsvirta & Åsbrink (1998) and others. This slow decay implies that the size of the returns is highly correlated, and volatility is a measure of the "average" size of those returns. Hence, persistent autocorrelation in the absolute returns is indicative of a smooth volatility process. This framework does not account for discontinuous structural breaks in volatility which may occur in practice (see Lamoureux & Lastrapes (1990), Hamilton & Susmel (1994), McConnell & Perez-Quiros (2000) for example), although our methodology could no doubt be adapted to better identify such changes. This remains a topic for future research, and a potential weakness of our proposed technique.

Many parametric models also support our smoothness assumption. Black & Scholes (1973) assume constant volatility, whereas the constant elasticity of variance model of Cox & Ross (1976) specifies volatility as a power function of the relatively smooth stock price process. Stationary generalised autoregressive heteroscedasticity (GARCH) models assume unconditional volatility is constant, and in certain cases allow conditional volatility to be a typically smooth process. Bollerslev & Mikkelson (1996) present a recently developed class of these models, designed to address "the apparent persistence of the estimated conditional variance processes" (Bollerslev & Mikkelson 1996, page 152). This long-memory is handled using fractional integration in which a shock to the conditional volatility estimate dies out at a slow rate in the future estimates.

The second key assumption is that the ϵ_t of (3.1) are independent, and have heavy-tailed distributions that are better approximated by a t-distribution (with a small number of degrees of freedom), than a Gaussian distribution. There are many studies (Fama (1965) being the first) that support the general heavy-tailed hypothesis, which would appear to be a ubiquitous feature of financial data. A number of candidate distributions have been proposed, of which the t-distribution is a common choice. See, for example, Blattberg & Gonedes (1974), Harvey et al. (1994), Hurst & Platen (1997), Liesenfeld & Jung (2000) and Barndorff-Nielsen & Shephard (2001) among many others. Typically the degrees of freedom ν of the t-distribution found in such studies range between 3 and 9. The t_ν distribution has infinite moments of order k when $k \geq \nu$, and so $\nu \geq 3$ ensures finite variance and $\nu \geq 5$ ensures finite kurtosis.

The commonly observed leptokurtosis in stock returns is not inconsistent with global parametric models for stock price processes. Volatility is often defined in a continuous time setting via a stochastic differential equation for stock price. The stock price process is generally adapted to Brownian motion, which has Gaussian increments. Under time-varying volatility, price returns, which are the increments in the log price process, can be leptokurtic (see Barndorff-Nielsen & Shephard (2001) for example). Thus, even if the driving stochastic process is assumed to have Gaussian increments, evolving volatility and heavy-tailed returns are theoretically linked.

In order to identify the volatility σ_t in (3.1), we require a final assumption that the ϵ_t have finite variance. Subject to this assumption, a global correction factor is computed from the sample variance of the original data standardised by the local scale estimates. This takes into account the fact that different scale estimators have different expected values for the same target distribution, as demonstrated in Table C.3. Although the heavy tailed distributions of the returns make the moving sample variance an unreliable *local* volatility estimate, particularly if the moving time window is small, we assume that the sample variance is a reliable estimator of the variance of the underlying heavy-tailed distribution in large samples of the order of the length of the data. In essence we trust that the extreme values of the heavy-tailed distribution will appear representatively in large samples where they will not distort the sample variance. However, in small samples of the order of the moving time window, such extreme observations will be over-represented when they occur and can severely distort the sample variance (as demonstrated in Chapter 2 for one-wild and slash data).

In conclusion, the minimal assumptions our volatility estimator will be based on appear reasonable, and have considerable support in both the empirical and theoretical literature.

3.2 Existing non-parametric methods

The methods we discuss in this section are general purpose techniques for estimating volatility. We describe them as non-parametric since they can be applied to data without specific modelling of the underlying stock price process. These methods typically have underpinning assumptions which are consistent with one or more parametric models; however the techniques are also typically used without direct reference to these assumptions (and indeed are used when these assumptions are deemed unreasonable).

3.2.1 Historical volatility estimation

A natural way to measure the slowly changing volatility σ_t in (3.1) is to take discretely sampled stock prices, form the returns, and estimate the standard deviation of these returns using a time series smoothing technique. Volatility estimates for real data formed on this (or any) basis will be difficult to appraise since the true volatility is unobservable; however, (3.1) and the notion of smoothness provide a useful framework.

A popular estimator of volatility is the historical volatility estimator. It has been prominent in empirical studies of stock returns at least since Officer (1973), and plays an important part of Figlewski (1997), where it is used to estimate "historical" and "realized" volatility as a basis for evaluation of volatility forecasting techniques.

Definition 3.1 (Historical volatility) *Historical volatility for a stock price series, with (mean-corrected) daily returns R_t is given by*

$$V_t = \sqrt{\frac{\sum_{j=-r}^{r} R_{t+j}^2}{2r}} \tag{3.2}$$

where $2r+1$ is the span of the estimator.

There are two important assumptions underlying use of this estimator, which is essentially a moving standard deviation. The first is that the volatility is roughly

constant over the length of the window, and this accords with the smoothness assumption we rely on in the derivation of our estimator. The second is that the sample standard deviation will be a reasonable estimator of scale for the observations in question. As we have seen in the analysis of Chapter 2, if the returns are not Gaussian, then the sample standard deviation may be highly inefficient, and as a result, (3.2) will provide poor estimates of volatility. We seek to address this second assumption in Section 3.3.

Figlewski (1997) computes (3.2) for the Standard and Poors 500 Index, with an annualisation factor. He notes that the volatility estimate is highly variable, and that it is unduly influenced by the October 1987 stock market crash. As the single extreme return of 19 October 1987, when the market lost 22.8% of its value, enters the 501 observation window (corresponding to $r = 250$ and roughly two years' data), it causes the volatility to instantaneously increase by 45% as shown in Figure 3.1. Exactly 500 trading days later, when the return is dropped from the estimation window, volatility falls by 23%. This effect can be controlled by using smoothness weights, so that as an observation moves from the end of the window (corresponding to terms $j = \pm r$ in (3.2)) to the central point of the estimation window (corresponding to the term $j = 0$) its weight increases. In this way, volatility would not change a great deal from one day to the next, but in the absence of robust fitting, it would still become very large close to the crash date. Using `loess`, described in Appendix A, with its tricube smoothness weights and no robustness properties, the volatility estimate reaches a maximum of 32.3% due to the disproportionate weight on the extreme return at the centre of the window.

3.2.2 Alternative non-parametric volatility estimation techniques

The historical volatility is usually calculated using returns based on consecutive closing prices (price at last trade). However, closing price is not the only variable relevant to volatility estimation that is available on a daily basis. In addition to closing price, opening price (price at first trade) and the daily high and low are also available, and these will contain useful information on the variability of the stock. These additional variables have been included in volatility estimators proposed by Parkinson (1980), Garman & Klass (1980), and Kunitomo (1992). These estimators

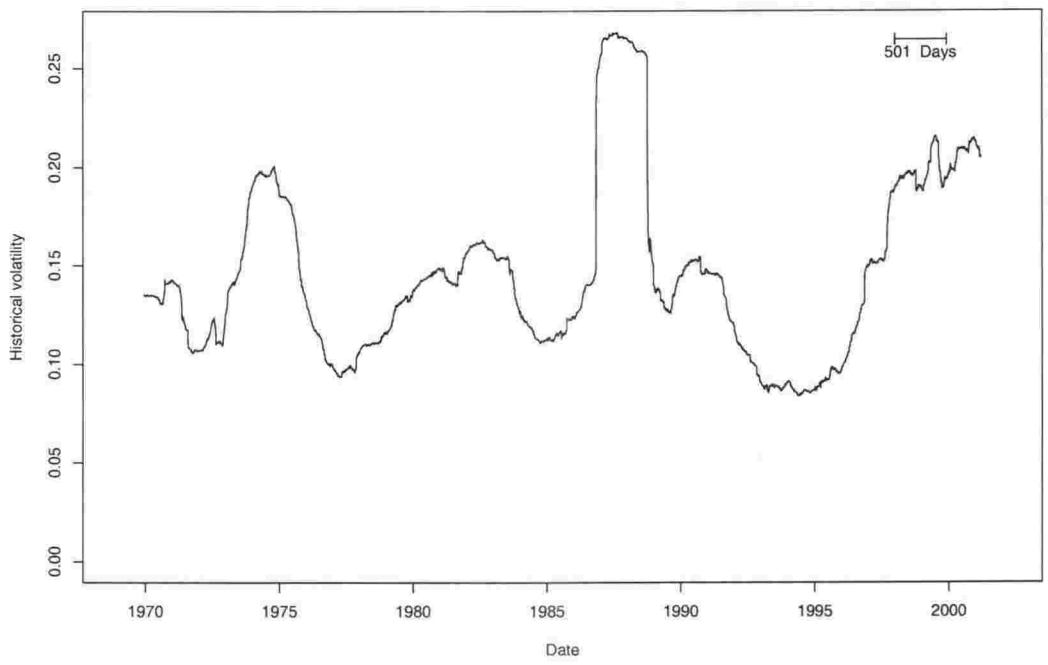

Figure 3.1. Historical volatility of the S&P 500 Index computed using (3.2) with $r = 250$. The figure is a reproduction of Figlewski's (1997) Figure II.2. Dates are indicated at the start of the respective years and the length of the smoothing window is shown.

were shown to be much more efficient estimators of constant volatility than the historical volatility (3.2) based exclusively on closing prices. However it is unclear how this efficiency will be affected by long-tailed data. Correction factors are provided for unbiasedness in the Gaussian situation, and time series estimation is not considered. For time-series estimates, the correction factor developed in Section 3.3.2 will be relevant. The impact of non-Gaussian errors on the estimators is likely to be similar to the effect on the sample standard deviation based estimators, and the gains in efficiency due to utilisation of additional market information may be lost due to lack of robustness.

In contrast to the symmetric smoothing window of the historical volatility estimator (3.2), the RiskMetrics software (JPMorgan 1996) uses simple exponential smoothing of the squared returns to estimate the volatility. The variance at time t is a weighted average of all past squared returns, with the weights decaying exponentially back through time. It uses the recursion

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)R_t^2$$

where the smoothing constant λ is chosen to be 0.94. This estimator will not be robust to extreme returns, however shocks to the estimated volatility will impact on future volatility estimates at an exponentially decaying rate.

Aït-Sahalia (1996) describes a non-parametric technique for estimating the volatility function of stationary interest rate data. He estimates the marginal density function of the spot interest rate series r_t using kernel density estimation, and derives an expression for volatility based on this estimated density, and an assumed drift function of the process. This technique is based on continuous-time specification for the interest rate, and the volatility function is estimated non-parametrically as a function of the rate rather than of time.

Anderson & Grier (1992) propose a non-parametric and robust definition of volatility, which is useful only when comparing the volatility of two series. Since it does not provide a numerical estimate of volatility, the estimate is neither useful for modelling stock price evolution, nor for valuing options on that stock.

The above non-parametric techniques, including historical volatility, are suitable techniques when the distribution of the innovations ϵ_t in (3.1) are approximately normal. In other situations, they are likely to be adversely affected by extreme returns, and consequently unsuitable generally. We address these concerns in the following section, where we utilise modern statistical techniques to provide a robust volatility estimator.

3.3 Robust volatility estimation

The volatility estimator of Anderson & Grier (1992) described above is robust, however it is not able to provide a time-series estimate of volatility. In contrast, Turner & Weigel (1992) robustly estimate volatility for the Standard and Poors 500 Index (S&P 500), and the Dow Jones Index. They use (3.2), and the volatility estimators of Parkinson (1980) and Garman & Klass (1980) which utilise high and low daily prices. In addition, volatility is estimated using a robust scale estimator: the interquartile range (IQR) defined in Definition 2.8. A disappointing aspect of this study is that the IQR-based estimates are not directly compared to the other estimators due to the bias effect shown in Table C.3. This method of estimation, and a correction that will allow direct comparison of any local volatility estimates, are discussed in Section 3.3.2.

In Chapter 2 the results of a simulation study to identify efficient robust scale estimators were described. The results included analysis of robust estimators examined by Lax (1985) in a similar study, but were also able to include two new single-pass

estimators (S_n and Q_n of Rousseeuw & Croux (1993)), both as scale estimators themselves and as auxiliary scale estimators for more complicated estimators. As well as the biweight A-estimator, which was the best performing estimator in the Lax study, estimators based on the t-distribution were introduced and these were shown to perform well for the three corner distributions: the normal, one-wild and slash.

The best performing estimator in the simulation described in Chapter 2 is the biweight A-estimator with auxiliary scale estimator Q_n (defined in (2.20)) and scaling constant $c = 11$. This estimator has the form

$$s_\psi(\mathbf{X}; c, Q_n) = \left[\frac{1}{n-1} \frac{\sum_{i=1}^n w(U_i)^2 (X_i - M)^2}{\left[\frac{1}{n}\sum_{i=1}^n \psi'(U_i)\right]^2} \right]^{\frac{1}{2}} \tag{3.3}$$

where M is the sample median,

$$U_i = \frac{X_i - M}{cQ_n}$$

is the standardised score, $w(x)$ is the biweight function

$$w(x) = \begin{cases} (1-x^2)^2 & |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

and $\psi(x) = xw(x)$. The average triefficiency of this estimator for $n = 20$ in the simulation study was 82.1% with average efficiency for normal data of 89.4%. Thus, under general conditions, we would expect this estimator to provide a reasonably efficient estimate of scale. Since volatility is a measure of the scale of daily returns, we can apply (3.3) in the same way as (3.2) to provide a moving volatility estimate.

Triefficiency was measured as the minimum of the efficiencies obtained in the three cases where the data follows the Gaussian, one-wild and slash distributions. These three "corner" distributions have varying degrees of heavy tail behaviour and are meant to delimit the situations met in practice. In particular, since the slash distribution has infinite variance and heavier tails than the target family of t_ν distributions ($\nu \geq 3$) that we have in mind here, we might expect higher efficiency (than the triefficiency) on application to financial data.

3.3.1 An alternative volatility estimator

We now develop an alternative estimator to (3.3) which is more closely tailored to the t-distribution. The results which follow are closely related to the development of the t-estimator in Section 2.3.3.

Let $X_1, \ldots, X_n$ denote n independent and identically distributed scaled t_ν random variables. The scaled t_ν variate follows a Gaussian compound scale model described in Definition 2.9, and can be written

$$X_t = \mu + \sigma \frac{Z_t}{\sqrt{S_t}}$$

for each t, where Z_t is a standard normal random variable and S_t is an independent chi-squared random variable with ν degrees of freedom, divided by $\nu - 2$. Under these conditions, and the additional restriction that $\nu \geq 3$, the scaled t_ν variate has mean μ and variance σ^2. In order to estimate μ and σ^2, we can apply Theorems 2.1 and 2.2 to obtain the maximum likelihood recursions.

Theorem 3.1 *The maximum likelihood estimators of location and scale for a random sample from the scaled Student's t-distribution with $\nu \geq 3$ degrees of freedom and variance σ^2, are found by iterating equations (2.8) and (2.9) with*

$$E_0(S_t|\mathbf{X}) = \frac{\nu+1}{\nu-2}\left(1 + \frac{(X_t - \hat{\mu}_0)^2}{(\nu-2)\hat{\sigma}_0^2}\right)^{-1}$$

where $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ are the previous estimates of μ and σ^2 respectively, and where ν is assumed known.

Proof Since S_t has cdf $P(S_t < s) = P(\chi^2 < s(\nu-2))$, its density is given by

$$f_{S_t}(s) = (\nu-2) f_{\chi^2_\nu}\big((\nu-2)s\big) = \frac{\nu-2}{\Gamma(\frac{\nu}{2})}\left(\frac{1}{2}\right)^{\frac{\nu}{2}} \big((\nu-2)s\big)^{\frac{\nu}{2}-1} e^{-\frac{1}{2}(\nu-2)s} \qquad (s>0)$$

for all t. Thus

$$\sqrt{s} f_S(s) \propto s^{\frac{1}{2}(\nu+1)-1} e^{-\frac{1}{2}(\nu-2)s}$$

which is in turn proportional to a gamma density function with parameters $\frac{1}{2}(\nu+1)$ and $\frac{1}{2}(\nu-2)$, and it follows from the proof to Theorem 2.7 that

$$\mathrm{E}_0(S_t|\mathbf{X}) = \frac{\nu+1}{\nu-2}\left(1 + \frac{(X_t - \hat{\mu}_0)^2}{(\nu-2)\hat{\sigma}_0^2}\right)^{-1}$$

as required. □

Application of Theorems 2.1 and 3.1 for sample data yields the maximum likelihood estimates of μ and σ^2 under the assumption that the sample is a random sample from the scaled t_ν distribution. For given degrees of freedom ν, the EM recursions will need to be iterated to obtain the maximum likelihood estimator of σ^2, and it is

the moving-average equivalent of this estimator that we choose to base our volatility estimation procedure on. Thus, we form moving windows as in (3.2), and estimate the scale in this window using Theorem 3.1 for each sub-sample of the returns data.

Other possible choices could be considered for the distribution of S_t that involve censoring to minimize the impact of outliers, mixed distributions with point mass at zero to account for sticky prices, and mixture distributions among other candidate distributions chosen to exemplify the target family of distributions under study. These remain topics for further research.

3.3.2 Local volatility estimation

We now consider the time series of daily returns R_t $(t = 1, \ldots, T)$ defined by (3.1) with heavy-tailed ϵ_t and where the R_t have been corrected for evolving level if appropriate. Natural time series estimators of the evolving volatility σ_t based on (3.3), Theorem 3.1 or more generally (2.9) are the finite moving averages of span $n = 2r + 1$ given by

$$\hat{\sigma}_t^2 = \sum_{j=-r}^{r} w_j q_{t+j} R_{t+j}^2 \tag{3.4}$$

where the smoothness weights w_j satisfy $\sum_{j=-r}^{r} w_j = 1$ and the robustness weights

$$q_t = Q\left(\frac{1}{2}\left(\frac{R_t}{s_t}\right)^2\right)$$

depend on a prior estimate s_t of σ_t. Here the function $Q(t)$ can be suitably defined for (3.3) or the estimator of Theorem 3.1 (refer to the proof of Theorem 2.7). In practice the estimates will involve iteration so that initial estimates of σ_t are successively refined. If (3.4) is based on (3.3) then s_t will be a moving cQ_n estimator, and only two iterations through the data are required, one to determine s_t and the other to determine the final estimate of σ_t. In the case of Theorem 3.1 one could iterate until approximate convergence, using the moving sample standard deviation (3.2) as the initial estimate of σ_t (a limited simulation study shows that a total of 4 iterations is usually sufficient), or use one iteration with s_t estimated by (3.3) yielding a total of 3 passes through the data.

We do not address methods for end-effect correction, and no volatility estimate is provided when all the $n = 2r+1$ observations needed for $\hat{\sigma}_t$ in (3.4) are not available, i.e. for $t = 1, \ldots, r$ and $t = T - r + 1, \ldots, T$. In the data analysis that follows, this

is accounted for by discarding the missing values: for the simulations we begin with returns series longer than the required volatility series, and for analysis of market return data we simply lose observations at either end of the series. Thus a series of length T provides a volatility series of length $T - 2r$. A simple alternative to omitting estimates at the ends, is to appeal to the assumption that the volatility is constant over the length of the window and use $\hat{\sigma}_t = \hat{\sigma}_{r+1}$ for $t < r+1$ and $\hat{\sigma}_t = \hat{\sigma}_{T-r}$ for $t > T - r$.

During the discussion of the robust scale estimation simulation of Chapter 2, it was noted that for data from any particular distribution, the scale estimators considered will estimate different factors of the distribution's scale parameter. This effect is highlighted in Table C.3. Thus, any attempt to estimate volatility robustly will require a correction to be made. However in general, this correction will depend on both the estimator in question, and the true distribution of the ϵ_t in (3.1). If a single historical estimate is provided along the lines of (3.2), the correction factor will need to be identified by theory or simulation, however, with local volatility estimation, a correction can be made using the returns R_t, and the time series volatility estimate $\hat{\sigma}_t$. This is formalised in the following definition.

Definition 3.2 (Local volatility estimator) *A local volatility estimator $\hat{\sigma}_t$ for the returns data $\{R_t\}$, $t = 1, \ldots, T$ has the following properties:*

1. *the local volatility estimator with span $n = 2r + 1$ is given by*

$$\hat{\sigma}_t = S(R_{t-r}, \ldots, R_{t+r})$$

   where $S(\mathbf{X})$ is a scale estimator;

2. *the standardised returns $\{(R_t - \hat{\mu}_t)/\hat{\sigma}_t\}$ $t = 1, \ldots, T$, have unit sample variance, where $\hat{\mu}_t$ is an estimate of the evolving mean return μ_t, as specified in (3.1).*

If the scale estimator used to provide the local volatility estimates has robustness properties, then the local volatility estimate will be a robust estimate.

The second condition of Definition 3.2 generally necessitates a correction to the volatility estimates obtained using a scale estimator. This correction is based on our assumption that the ϵ_t are zero mean, unit variance random variables. We assume that although not a robust or reliable estimator in small samples, the sample variance

can be expected to provide a reliable estimate of the variance for very large samples. As noted in Section 3.1, the extreme values of the heavy-tailed distribution of ϵ_t are assumed to appear representatively in large samples of the order of the series length T, but will be over-represented when they occur in small samples of the order of the span $n = 2r + 1$ of the moving estimation window.

We assume our volatility estimators typically estimate $\sqrt{\sigma_t^2/\tau}$ where τ is a positive constant and $\tau \neq 1$. To correct for this bias we multiply through (3.4) by $\hat{\tau}$, where

$$\hat{\tau} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{R_t - \hat{\mu}_t}{\hat{\sigma}_t}\right)^2 \tag{3.5}$$

where $\hat{\mu}_t$ is an estimate of the evolving mean return μ_t, as specified in (3.1). This estimator is just the sample variance of the scale adjusted returns, and would have enabled Turner & Weigel (1992) to compare the volatility estimate based on the interquartile range (IQR) to their other estimates of historical volatility. In their study of the S&P 500 and Dow Jones indices, Turner & Weigel calculated the sample standard deviation and interquartile range of daily returns in each of the calendar years from 1928 through to 1989. In addition, the volatility estimators of Parkinson (1980) and Garman & Klass (1980), which utilise high and low daily prices (as well as closing price in the latter case), were also computed. These two estimators feature correction factors to ensure unbiasedness for the parameter σ under the assumption that the daily returns are normal and locally have constant variance σ^2. For Gaussian X, $\mathrm{E}(\mathrm{IQR}(X)) = 1.3490\sigma$, and so the IQR-based volatility estimates should be on average 1.3490 times those from the other estimators, and hence they were not directly comparable in the graphs given by Turner & Weigel (a larger order difference arises because the non-robust estimates are annualised, while the IQR-based estimates are not). Despite this, it is clear from the graphs given by Turner & Weigel that the two have roughly the same character.

Definition 3.3 (Iterated t-volatility estimator) *The iterated t-volatility estimator, with degrees of freedom parameter $\nu = 5$, is a local volatility estimator satistying the conditions of Definition 3.2, and is given by Theorem 3.1 for each window. Iteration may be to convergence of the local estimates, or alternatively, a specified number of iterations may be performed.*

We selected $\nu = 5$ for the iterated t-volatility estimator since the t_5 distribution has heaviest tails among the t-distributions with finite variance and kurtosis, and it

reflects the intermediate case identified in empirical studies of daily returns. Ironically, this choice of degrees of freedom was not reported in the simulation results of Chapter 2. Nonetheless, we can see from the results in Table 2.12 and Figure 2.11 that the fully iterated estimator with $\nu = 5$ would have triefficiency of approximately 42% for samples of 20 observations (due to a low slash efficiency), but normal and one-wild efficiencies in excess of 80%.

In the following section, the iterated t-volatility estimator is compared to other local volatility estimators, for various underlying distributions for the ϵ_t in (3.1).

3.4 Simulations and data analysis

Using simulated and real data, we now consider the relative performances of the local volatility estimators based on the standard deviation, MAD, biweight A-estimator with Q_n and $c = 11$, and iterated t-volatility estimator respectively. In all cases the iterated t-volatility estimator was initialised by the local sample standard deviation and iterated a further three times to produce a final estimate.

3.4.1 Simulation results

For the simulation study we simulated 270 returns from the scaled t_ν distribution with unit variance and with $\nu = 3, 5, 9$, and from the standard normal distribution (i.e., t_∞), both to represent varying degrees of heavy tailed behaviour and to be appropriate for financial daily returns data. The series length was chosen to roughly represent a calendar year of trading days allowing for end effects. Our estimators were based on moving windows of span $n = 21$ with uniform weights $w_j = 1/21$. The latter were selected since our concern at this stage was with the precision of the estimators rather than their smoothness. The impact of the smoothness weights w_j on the properties of the estimators, among other issues, remains to be investigated. The 270 scaled t_ν returns were then multiplied by the smooth volatility function σ_t where

$$\sigma_t = 3e^{\frac{1}{2}\sin(\pi t/125)}.$$

The four estimates of σ_t were calculated and scaled using (3.5) so that the standardised returns had unit sample variance. They were assessed by the mean absolute proportionate error of the squared volatility (MAPE $= \frac{1}{T}\sum |\sigma_t^2 - \hat{\sigma}_t^2| / \sigma_t^2$) for each

Estimator	t_3	t_5	t_9	t_∞
Standard deviation	0.803	0.439	0.346	0.280
Median absolute deviation	0.629	0.509	0.485	0.467
Biweight A-estimator with Q_n and $c = 11$	0.554	0.394	0.349	0.306
Iterated t-volatility estimator with $\nu = 5$	0.526	0.340	0.297	0.263

Table 3.1. The average mean absolute proportionate error (MAPE) of four local volatility estimators estimating a smoothly varying volatility function over moving windows of span 21. The average is over 10000 simulated series, each with an individual MAPE. The simulated returns have scaled t_ν-distributions with $\nu = 3, 5, 9$ and ∞, the latter case being a normal distribution.

estimator, computed over the $T = 250$ volatility estimates available. These statistics were then averaged over 10000 independent realisations of the time series, for each of the four distributions, to yield the results in Table 3.1.

The results are self evident. The iterated t-volatility estimator performed best, even in the cases where the underlying distribution was not the t_5 distribution. As expected, the volatility estimator based on the moving standard deviation performed reasonably well for $\nu = 9$ and ∞, but its performance deteriorated as ν decreased. The biweight A-estimator performed reasonably well in all cases, and consistent with the results of Chapter 2, use of this more advanced estimator results in better estimates than the MAD. It might be expected that the A-estimator would have a comparable or better performance than the iterated t-volatility estimator for heavy tailed data not well-approximated by a t-distribution. However, aside from what can be inferred from the results of Chapter 2, this has not yet been verified.

The MAPEs themselves are shown in Figure 3.2. We see that not only is the average MAPE lower for the iterated t-volatility estimator in each sampling situation, but also the MAPEs in each case are generally less variable. For both the t_9 and normal data, the moving standard deviation, A-estimator, and iterated t-volatility estimator are all of very similar quality, with the distributions for the normal data being slightly less variable for all estimators.

We also consider the performance of the iterated t-volatility estimator through time. Figure 3.3 shows the proportionate errors $(\hat{\sigma}_t^2 - \sigma_t^2)/\sigma_t^2$ for $t = 10, 30, \ldots, 250$ for the 10000 simulated series. The function σ_t^2 is also shown for reference, with scale given by the right-hand axis. We see that there is generally a negative bias, and that this bias is largest when σ_t^2 has the greatest curvature. As the iterated t-volatility estimator smoothes the squared returns, the high degree of non-linearity over the smoothing window of the volatility function around $t = 50$ induces this negative

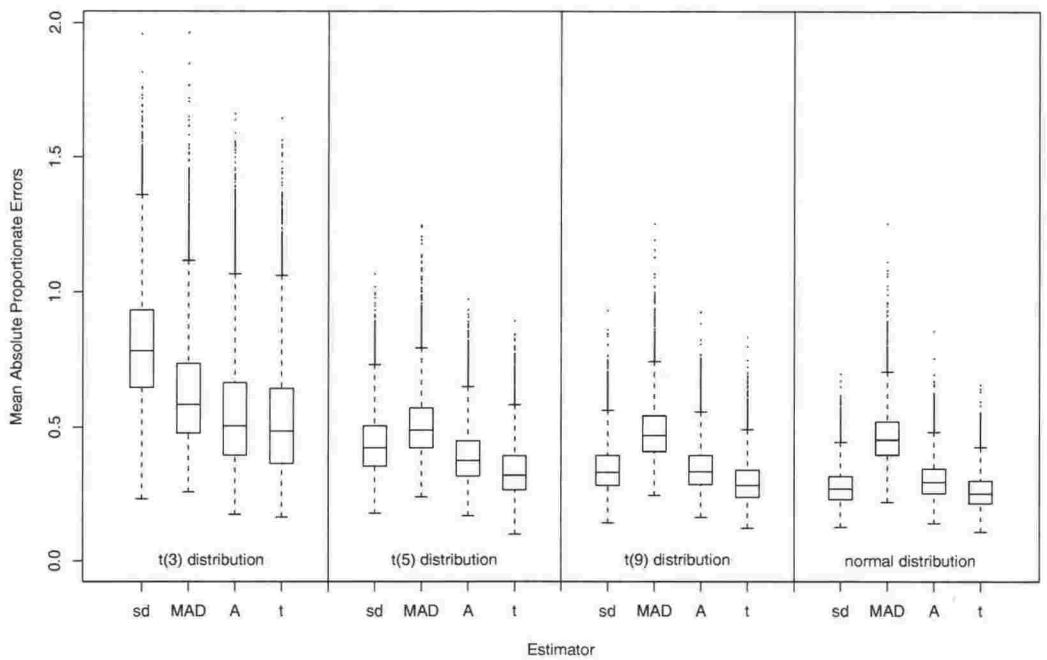

Figure 3.2. 10000 realisations of the mean average proportionate errors for each estimator, with sd denoting the moving standard deviation, MAD the moving median absolute deviation, A the moving A-estimator with Q_n and $c = 11$, and t the moving t-volatility estimator for series of 250 returns. The four blocks represent simulated returns with scaled t_ν-distributions with $\nu = 3, 5, 9, \infty$ respectively, where the latter case is a normal distribution.

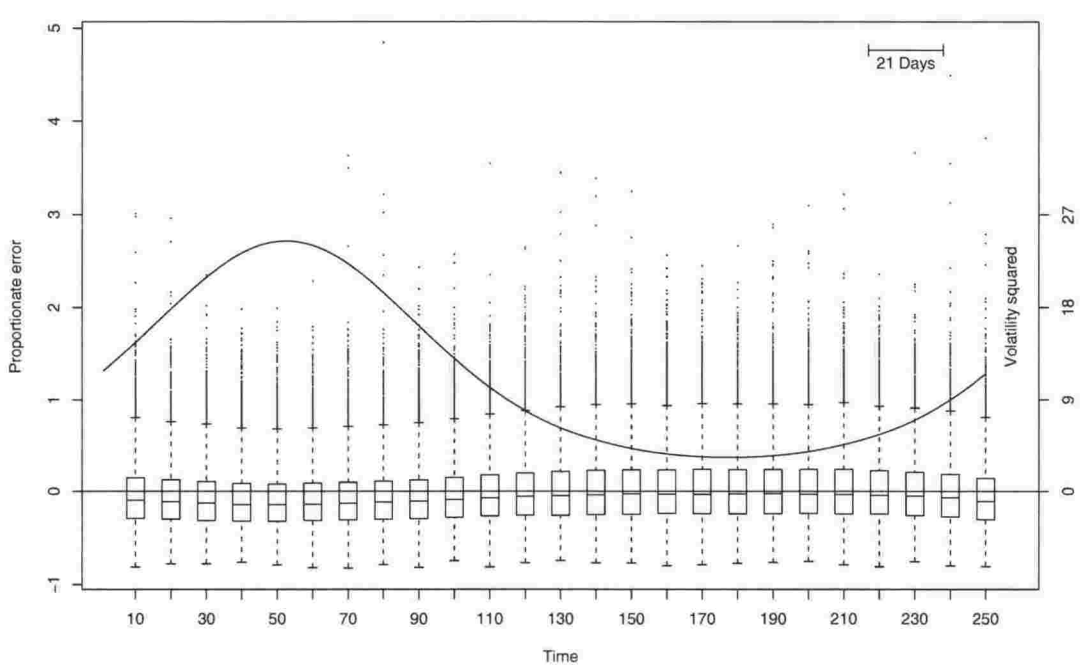

Figure 3.3. Boxplots showing the proportionate bias $(\hat{\sigma}_t^2 - \sigma_t^2)/\sigma_t^2$ through time, for 10000 simulated series of length 250. The simulated returns have a scaled t_5-distribution, and the function σ_t^2 is shown for reference, with magnitude given by the right-hand axis.

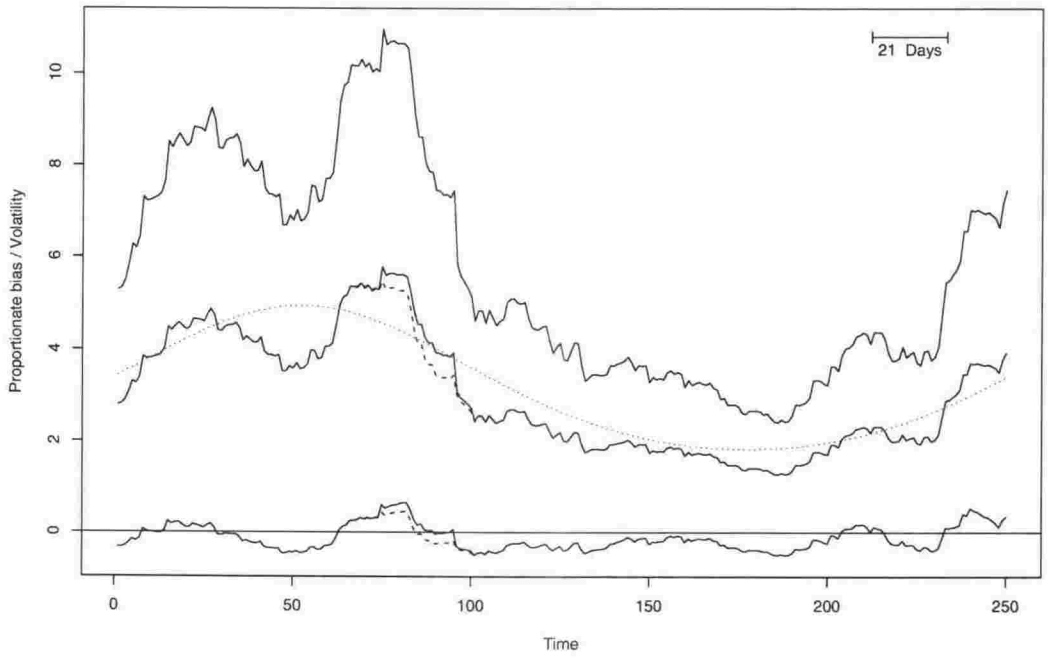

Figure 3.4. Volatility estimates, and their proportionate biases, for simulated t_5 returns. Five series are shown, along with the true underlying volatility process (using the dotted line). The upper estimate is the iterated t-estimate of Definition 3.3. The two estimates moving around the volatility function are: using the dashed line, the volatility estimate obatined by truncating a single value at $t = 75$, and using the solid line, the volatility estimate for the original data, rescaled using the factor from the truncated series. The estimates moving around the horizontal line at zero are the proportionate biases, $(\hat{\sigma}_t^2 - \sigma_t^2)/\sigma_t^2$ corresponding to the two lower volatility estimates.

bias. This implies that the choice of window width for the volatility function used is too high, and this is particularly pronounced when the volatility function is highly non-linear.

We also note that in some cases, the proportionate bias is very large. Examination of an offending series shows that this is related to the finite-sample bias correction (3.5), and occurs when a particularly extreme observation is realised in the series, and $\text{var}(R_t/\hat{\sigma}_t)$ is inflated as a result. In order to analyse the extent of this effect, the simulations were repeated for longer series ($T = 1000$ observations long). Comparison with Figure 3.3 shows that generally the bias distributions are very similar; however, some of the more extreme biases have been eliminated. This is due to the improved efficiency of the sample variance used in (3.5) for these series.

Figure 3.4 examines further the volatility estimate which has the largest proportionate bias at $t = 80$ in Figure 3.3 of approximately 5. Three volatility estimates are shown: the upper-most is the volatility estimate analysed in Figure 3.3, and this is clearly a poor estimate of the true volatility function. The second and third estimates differ only around $t = 75$, where a single very extreme observation (a standardised t_5 variate of -27.736, an observation that is expected to occur only once in every 880,000 observations) was present in the series. The solid curve is the volatility estimate obtained from the raw data, but rescaled using the second volatility estimate's correction factor. As a result, each estimate is $1/1.679$ times the original estimate. The dashed volatility estimate is based on the original series, but with the value -27.736 replaced by -12 (chosen so that it is still the most extreme value in the series). Because of the way they have been rescaled, we see that these volatility estimates are identical when the outlier is not included in the smoothing window, and both are reasonable estimates of the true volatility function, shown by the dotted line. Also shown in the plot is the proportionate error estimate $(\hat{\sigma}_t^2 - \sigma_t^2)/\sigma_t^2$ for each of the volatility estimates. These series oscillate around the horizontal line at zero, with the dotted line showing the difference corresponding to the truncation. These errors range between -0.54 and 0.52, clearly an improvement on the original, which had a range of 0.69 to 4.93.

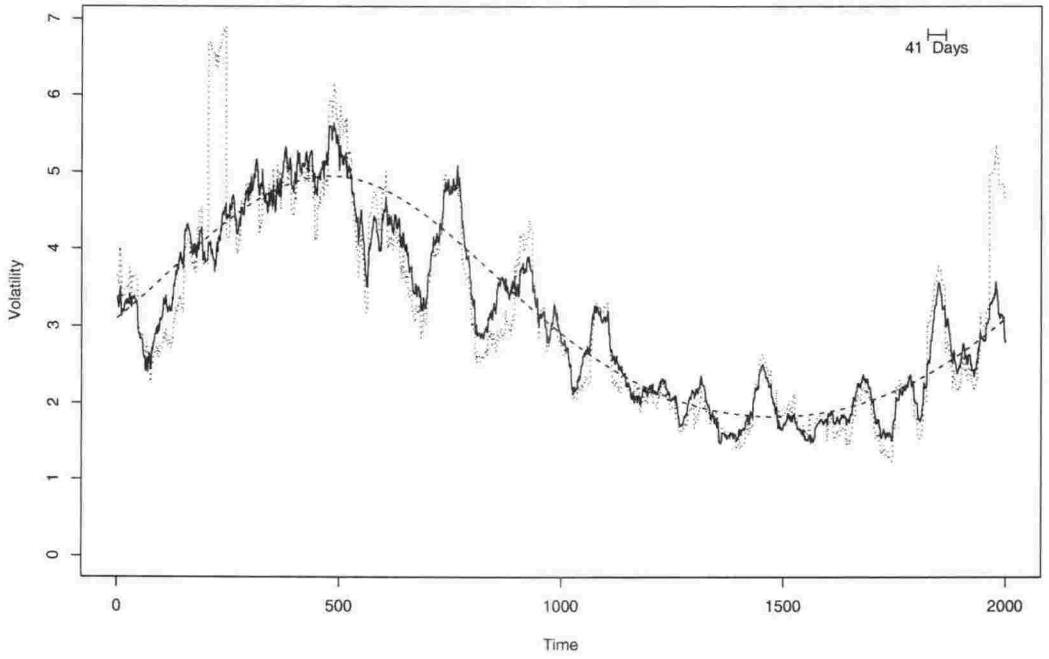

Figure 3.5. Volatility estimates for 2000 simulated t_5 returns with evolving volatility given by the dashed line. The solid estimates are given by the iterated t-volatility estimator in Definition 3.3. The dotted estimates are computed using (3.2), and have no robustness properties.

3.4.2 A simulated return series from the t_5 distribution

Further insight into the iterated t-volatility estimator, given by Definition 3.3, is gained by analysing a single simulated series. In this case, we generate 2040 returns according to (3.1), with $\mu_t = 0$,

$$\sigma_t = 3e^{\frac{1}{2}\sin(\pi t/1000)}$$

as before, so that the entire series has a single cycle of this volatility function, and ϵ_t is drawn independently from the t_5 distribution. A window length of 41 observations will be used for the analysis, and so the volatility estimate will be a series 2000 long. We use this series to check that the procedure is correctly identifying the evolving volatility σ_t, so that estimates of the innovations ϵ_t in (3.1) can safely be used to examine the underlying distribution of returns.

The estimated volatility using the iterated t-volatility estimator, and the traditional historical volatility estimator (the moving standard deviation (3.2)), both with a smoothing window of 41 observations, are shown in Figure 3.5. The estimates are very similar except in instances where extreme values are present in the smoothing window, at which time the non-robust estimate has large departures from the true volatility function. The estimators shown in Figure 3.5 accord with the general

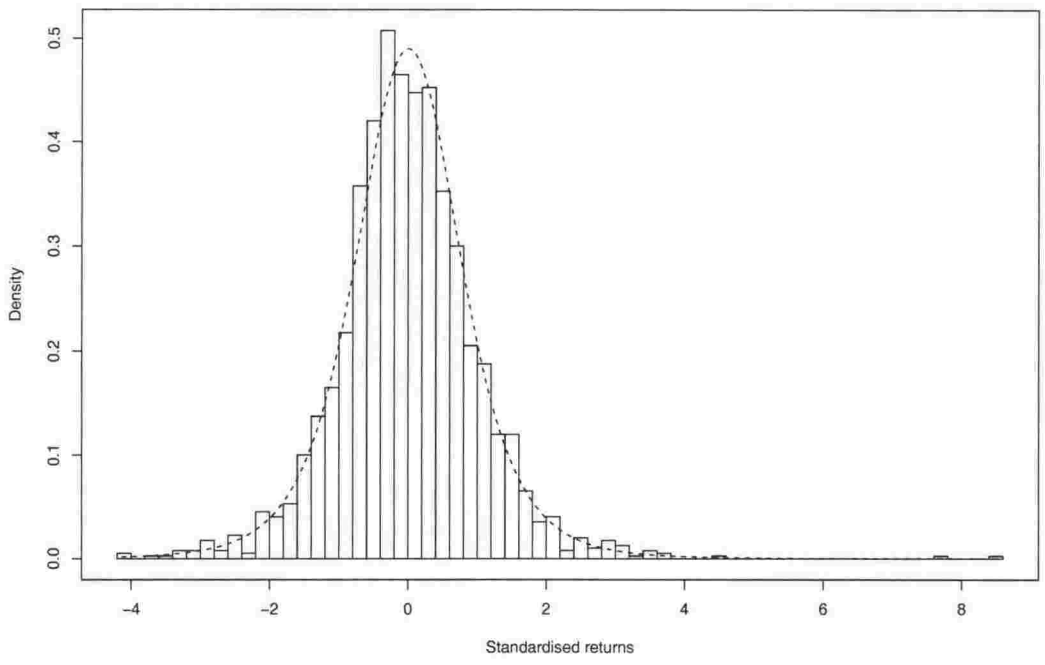

Figure 3.6. Standardised returns for the simulated t_5 returns analysed in Figure 3.5, along with the density function for the scaled t_5 distribution. The returns are standardised using the iterated t-volatility estimate in Figure 3.5.

performance of the iterated t-volatility estimator and the moving standard deviation in the more extensive study described in Table 3.1. In particular, for t_5 distributed data, the mean absolute proportionate error of the moving standard deviation is much greater than that of the iterated t-volatility estimator.

The volatility estimate provided by the iterated t-volatility estimator is used to standardise the returns. A mean of zero is assumed, and the innovations in (3.1) are estimated by $\hat{\epsilon}_t = R_t/\hat{\sigma}_t$. The sample distribution of these standardised returns is shown in Figure 3.6, and is compared to the density function of the scaled t_5 distribution. As we might expect, the two distributions match very well, and we conclude that $\hat{\sigma}_t$ is a reliable estimate of σ_t, and that the distribution of $\hat{\epsilon}_t$ is a reliable estimate of the true underlying distribution of ϵ_t.

Finally, we consider the sample autocorrelation function (ACF) of the absolute returns $|R_t|$, and of the absolute standardised returns $|\hat{\epsilon}_t| = |R_t|/\hat{\sigma}_t$. Since the ϵ_t are independent, the returns are also independent, and it follows that the absolute returns are independent. However, because of the evolving volatility σ_t, the absolute returns exhibit significant autocorrelation for very many lags, as seen in the upper plot of Figure 3.7. We would hope that a reliable estimate of volatility would account for this autocorrelation behaviour. Examining the ACF of the standardised

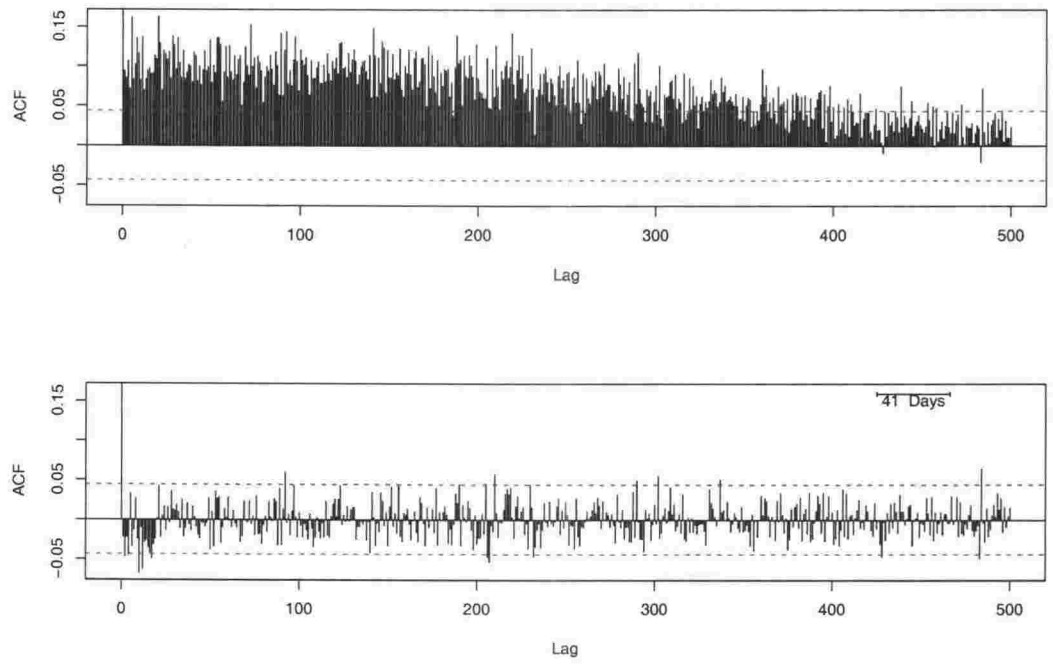

Figure 3.7. The estimated autocorrelation functions for the absolute simulated t_5 returns analysed in Figure 3.5, and for the absolute standardised returns. The top plot is for $|R_t|$, and the lower is for $|R_t|/\hat{\sigma}_t$ where $\hat{\sigma}_t$ is the volatility given by the iterated t-volatility estimator, shown in Figure 3.5. Approximate 95% confidence intervals for the autocorrelation estimates are shown. The two plots are on the same scale, and only $\rho_0 = 1$ is not shown.

returns, shown in the lower plot in Figure 3.5, we see that with the exception of some lags less than the smoothing window length of 41 observations, the autocorrelation has been almost completely removed from the absolute returns. This shows that the volatility estimation procedure we are promoting is correctly identifying the σ_t component in (3.1).

3.4.3 The S&P 500 data

The S&P 500 Index has been a much studied financial time series. The volatility of this series for the period 1969-2001 inclusive, calculated using (3.2) with a 501 day smoothing window, was shown in Figure 3.1. The indication there was that the historical volatility estimator based on a moving standard deviation was not resistant to the "crash" of October 1987. We examine this series further using the iterated t-volatility estimator of Definition 3.3.

Volatility is estimated for the S&P 500 using a window of 125 days (approximately half a trading year) in (3.2), using a moving interquartile range, and using the iterated t-volatility estimator. All estimates are corrected using (3.5) in order to

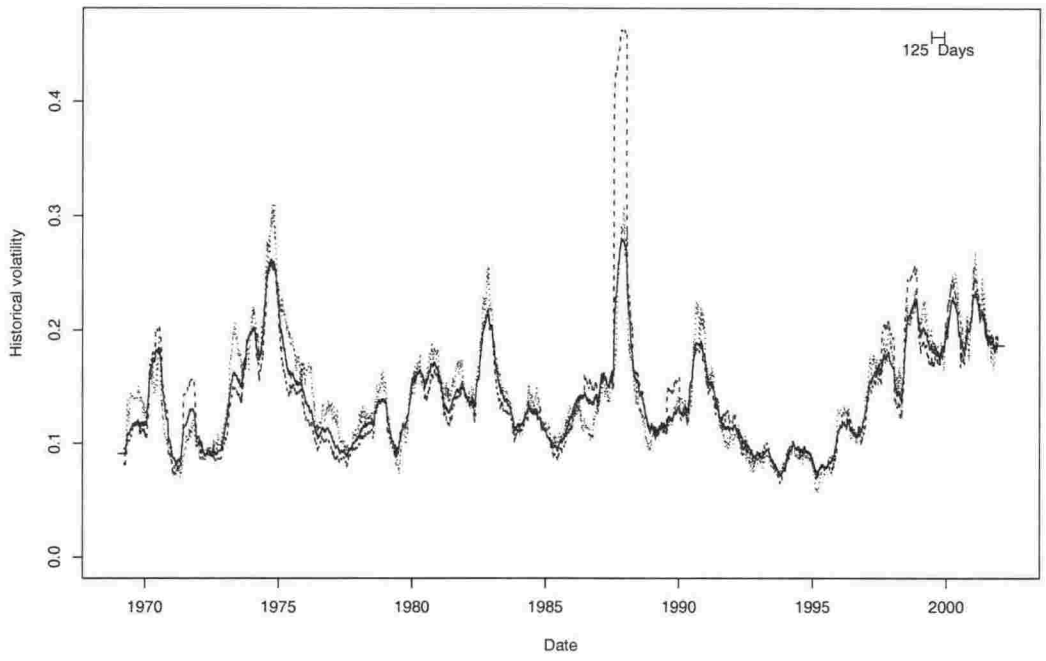

Figure 3.8. Robust volatility estimation for the S&P 500 index. The dashed estimate is computed using (3.2) and has no robustness properties. The solid line is based on the iterated t-volatility estimator with $\nu = 5$, and the dotted line is based on a moving interquartile range. All estimates have been corrected using (3.5) in order to satisfy Definition 3.2. A smoothing window of 125 days has been used throughout.

satisfy Definition 3.2, and are shown in Figure 3.8. There are 8652 returns in the series, and the window length for this plot is chosen to facilitate comparison between the estimates.

The estimates shown have many interesting features. Firstly, we note that the data used are identical to those used for the replication of Figlewski's (1997) historical volatility shown in Figure 3.1; however in this case a smaller smoothing window is used. We see the influence of the October 1987 stock market crash remains in the historical volatility estimate; however its overall impact on volatility is lessened due to the shorter window. Apart from the period centred around October 1987, the three volatility estimates are largely similar in nature. Consistent with the results of Chapter 2, the interquartile range, though robust, is not very efficient, and the volatility estimate it provides is less smooth than either of the other two estimates. The iterated t-volatility estimator is generally closer to the historical volatility estimate than the IQR-based estimate, except around the 1987 crash. Consistent with Turner & Weigel (1992), the IQR-based measure finds the 1987 period to be less volatile than the period from mid-1974 to mid-1975. The t-volatility estimator generally lies between the two alternatives, and although the three estimates are often

difficult to distinguish between, the non-robust estimate is clearly too high due to one-off returns in 1972, 1987, 1989 and 1998.

Standardising the S&P 500 returns using the volatility based on the iterated t-volatility estimator, but with a smaller smoothing window of 41 days, we can examine the distribution of these returns. The standardised returns are formed with (3.1) in mind, with $\hat{\mu}_t$ given by `loess` with a smoothing window of approximately 175 days (2% of the observations), and their distribution is shown in Figure 3.9. Two large negative standardised returns, namely -13.31 on 19/10/1987 and -8.05 on 13/10/1989, are omitted from the histogram. The density functions of the standard normal and the scaled t_5 distribution are superimposed on the histogram. While neither of these is a very good description of the tails of the distribution, the t_5 does a much better job of describing the centre of the distribution, both supporting the use of the iterated t-volatility estimator for this data, and strongly suggesting that the traditional historical volatility estimate is inappropriate. The effect of this non-normality on the historical volatility estimate is clear both from Figure 3.8 and the results of Chapter 2, and it is also likely to cause problems in estimators using daily price extremes (Parkinson 1980, Garman & Klass 1980, Kunitomo 1992).

Ding & Granger (1996) analyse the returns of the S&P 500 Index for volatility persistence, which is embodied in the slow decay of the sample autocorrelation function (ACF) of the absolute (or alternatively, squared) returns. Using the decomposition (3.1), we would want this feature to be explained by σ_t, and leave both ϵ_t and $|\epsilon_t|$ uncorrelated through time. In order to estimate ϵ_t, we standardise R_t using a slowly evolving mean $\hat{\mu}_t$ provided by `loess`, and a volatility estimate provided by the iterated t-volatility estimator, to obtain the standardised returns shown in Figure 3.9. These standardised returns are referred to as *rescaled returns* by Taylor (1986), who shows that they facilitate more precise autocorrelation estimates than the daily returns, due to their relatively constant scale.

The estimated ACF plots of the absolute returns and the absolute standardised returns are shown in Figure 3.10. In the top plot, we see that the absolute returns exhibit significant autocorrelation for very many lags, as described by Ding & Granger (1996). After standardising, we see that the autocorrelation is almost entirely accounted for by $\hat{\sigma}_t$. Almost all the significant autocorrelation estimates are at lags less than the smoothing window length of 21 daily observations. A similar phenomenon was seen in Figure 3.7 for the simulated data.

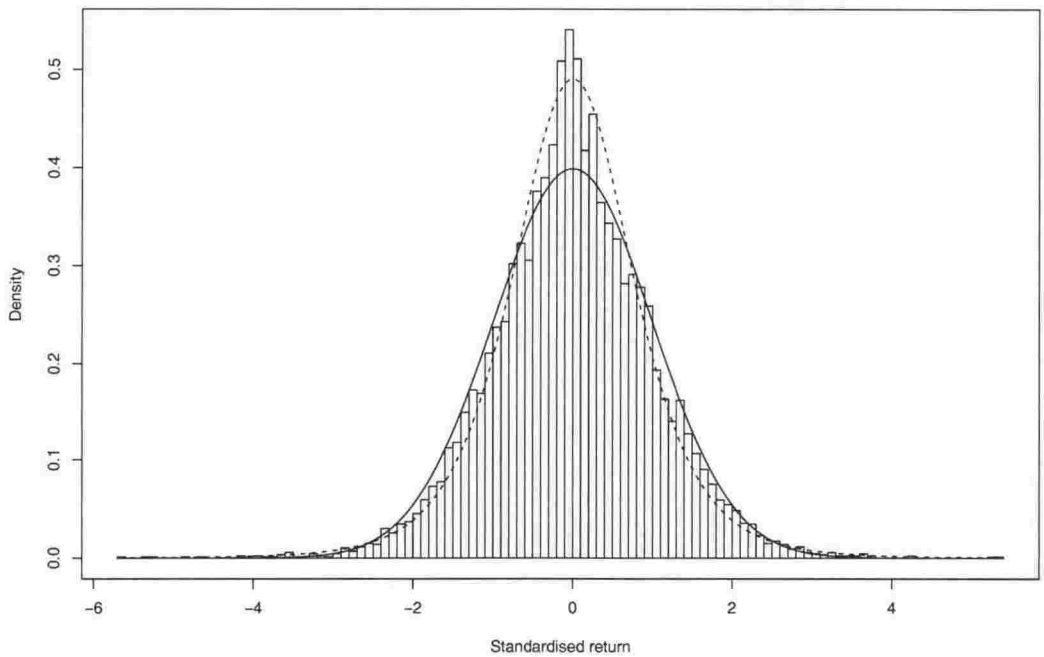

Figure 3.9. Standardised returns for the S&P 500 index along with the density functions for the standard normal (the solid curve), and the scaled t_5 distribution. The returns are standardised using a location estimate provided by `loess` and the t-based volatility estimate in Figure 3.8. Two standardised returns have been omitted from the lower tail: -13.31 on 19/10/1987 and -8.05 on 13/10/1989.

The combined evidence of Figures 3.9 and 3.10, and the poor performance of the historical volatility in Figure 3.8, reassure us that the iterated t-volatility estimator is a definite improvement on existing methods, and the resulting volatility estimates are excellent estimates of the underlying volatility process σ_t.

3.4.4 Individual Australian stocks

In this section, we provide a brief analysis of the volatility of two individual stocks. Coca-Cola Amatil Ltd (CCL) and The Broken Hill Proprietary Company Ltd (BHP) are among the largest and most actively traded companies listed on the Australian Stock Exchange. Daily closing price data for these stocks, for the 500 trading days preceding 1 September 2000, were analysed. A time series plot of the daily returns for each stock features periods of low volatility and periods of high volatility, and a small number of extreme returns. This latter phenomenon is more evident in the CCL returns.

Calculation of evolving volatility for CCL using the moving standard deviation (3.2) produces volatility estimates which are badly affected not only by the large returns,

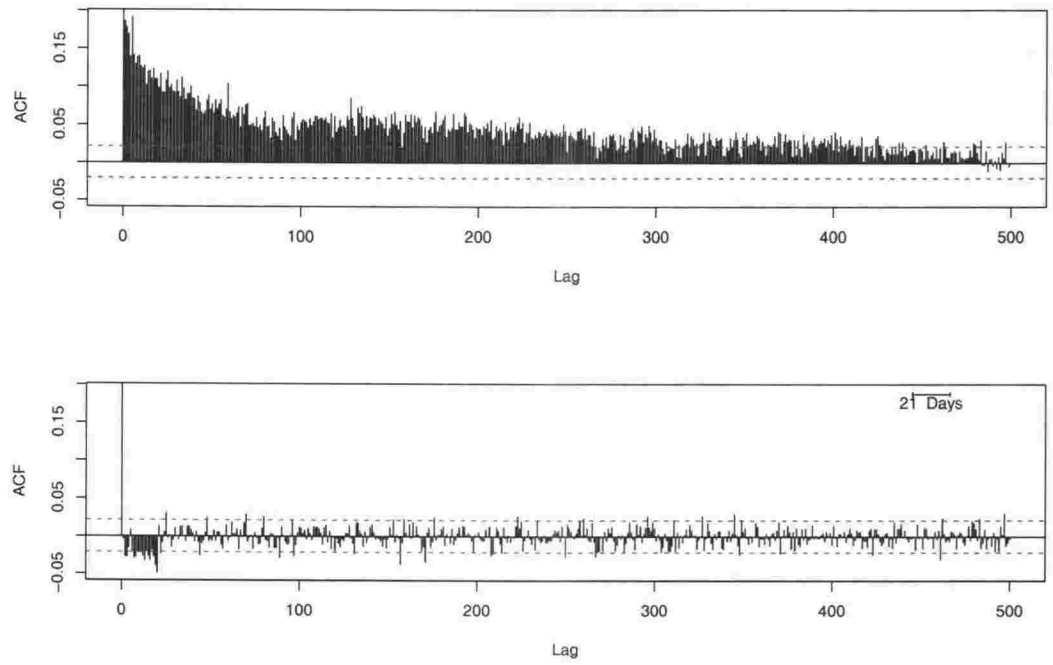

Figure 3.10. The estimated autocorrelation functions for the absolute S&P 500 returns analysed in Figure 3.8, and for the absolute standardised returns. The top plot is for $|R_t|$, and the lower is for $|R_t - \hat{\mu}_t| / \hat{\sigma}_t$ where $\hat{\mu}_t$ is estimated using `loess`, and $\hat{\sigma}_t$ is the volatility given by the iterated t-volatility estimator. Approximate 95% confidence intervals for the autocorrelation estimates are shown. The two plots are on the same scale, and only $\rho_0 = 1$ is not shown.

but also by the many small returns. The resulting fluctuations in the volatility estimates gives a distribution of standardised returns that is not well approximated by the Gaussian distribution since it has a sharp peak and values outside four standard deviations from the mean. Estimating scale using the iterated t-volatility estimator also results in standardised returns that are not well approximated by the normal distribution. However the distribution of these standardised returns has a smoother peak (the generally lower volatility estimates do not bring so many returns close to zero), and a distribution that is reasonably well approximated by a t_ν distribution with $\nu = 5$. On this basis, parametric volatility estimation based on the t_5 distribution should improve the quality of the volatility estimates.

A plot against time of the three estimates of evolving volatility based on the standard deviation, A-estimator and iterated t-volatility estimator for a smoothing window of $n = 41$ days is given in Figure 3.11 and shows that the iterated t-volatility estimator is the most stable. It is clear that the iterated t-volatility estimator typically adopts a compromise position between the standard deviation and the A-estimator, but closer to the A-estimator. The impact of extreme returns is clearly evident on the standard deviation and the A-estimator often appears to discount such returns

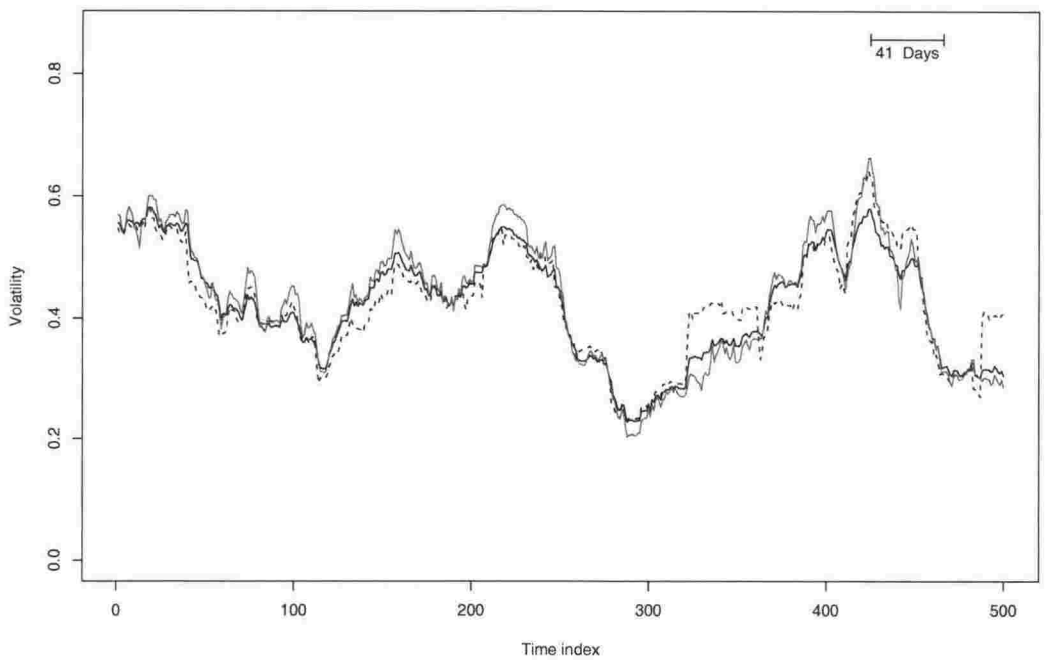

Figure 3.11. Absolute (annualised) returns and volatility estimates for CCL for the 500 trading days preceding 1 September 2000. The volatility estimates are: moving standard deviation shown by the dotted line, moving A-estimator with Q_n and $c = 11$ shown in dark grey, and moving t_5 estimator shown by the solid line. The largest 5% of the absolute returns (exceeding 5.5% in one day) are not shown in the plot area.

too heavily. The superior performance of the iterated t-volatility estimator is to be expected given the results of the simulation study reported in Table 3.1 and the fact that the distribution of standardised returns was well-approximated by a t-distribution.

In contrast, the distribution of the standardised returns for BHP *is* well approximated by a normal distribution. The tails of the sample distribution decay quickly and all observations are within 3 standard deviations of the mean. While the use of the moving standard deviation is appropriate for this data, the other two volatility estimators give almost identical volatility estimates as shown in Figure 3.12. Thus the iterated t-volatility estimator and A-estimator retain high efficiency in this situation also. Indeed, by using the iterated t-volatility estimator generally, it seems that we benefit in the case of long-tailed returns, and maintain high efficiency with well-behaved data.

The distributions of standardised returns for the CCL and BHP data are shown in Figure 3.13. In each case, the returns have been mean-corrected using `loess`, and standardised using the iterated t-volatility estimator with $\nu = 5$. Figures 3.11 and 3.12 showed quite different character in the robust and non-robust volatility esti-

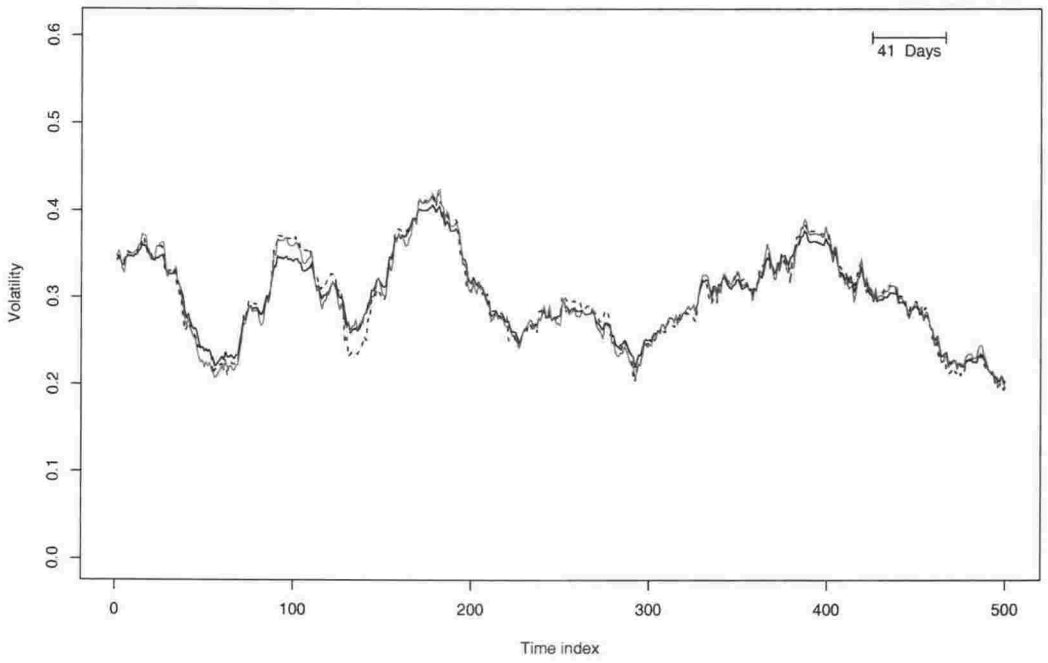

Figure 3.12. Absolute (annualised) returns and volatility estimates for BHP for the 500 trading days preceding 1 September 2000. The volatility estimates are: moving standard deviation shown by the dotted line, moving A-estimator with Q_n and $c = 11$ shown in dark grey, and moving t_5 estimator shown by the solid line. The largest 5% of the absolute returns (exceeding 3.8% in one day) are not shown in the plot area, and the plot is not in the same scale as that in Figure 3.11.

mates and these features are confirmed in the histograms. In particular, we see that the CCL returns are better described by the scaled t_5 distribution than the normal, both at the mode, and in the tails. Use of the iterated t-volatility estimator is thus justified, and the obvious differences between the volatility estimates confirms the unsuitability of the traditional estimator. For the BHP data, the three volatility estimates were very similar, and we find the standardised returns are well approximated by the normal distribution. While we note that this lends support to the traditional estimator of historical volatility for this series, we must also point out that the iterated t-volatility estimator provided a very similar series of estimates.

The estimated ACF plots of the absolute returns and the absolute standardised returns for both CCL and BHP are given in Figure 3.14. Unlike the equivalent plots for the simulated series and for the S&P 500 data, the absolute returns for these two series show very little autocorrelation. This is particularly true for BHP, and while this may be affected by the relatively small number of observations (500 compared to 2000 and 8652 for the simulated series and the S&P 500 respectively), we see in Figure 3.12 relatively constant volatility. Some colour has been removed from the ACF of the absolute CCL returns, however we see the same phenomenon as earlier,

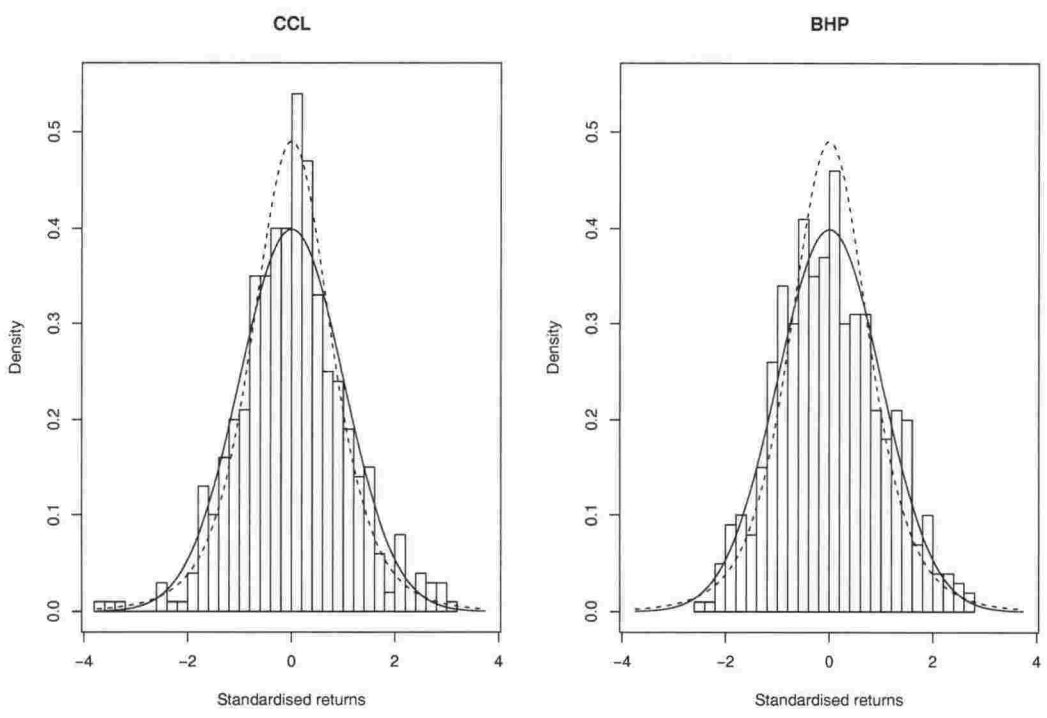

Figure 3.13. The distributions of the standardised returns for CCL and BHP. The daily returns are standardised using a mean from `loess`, and volatility provided by the A-estimator with Q_n and $c = 11$, as shown in Figures 3.11 and 3.12 respectively. The two plots have identical scales, and feature the standardised returns distribution for CCL on the left, and BHP on the right. Both distributions have density functions superimposed for the scaled t_5 (dotted) and the standard normal (solid).

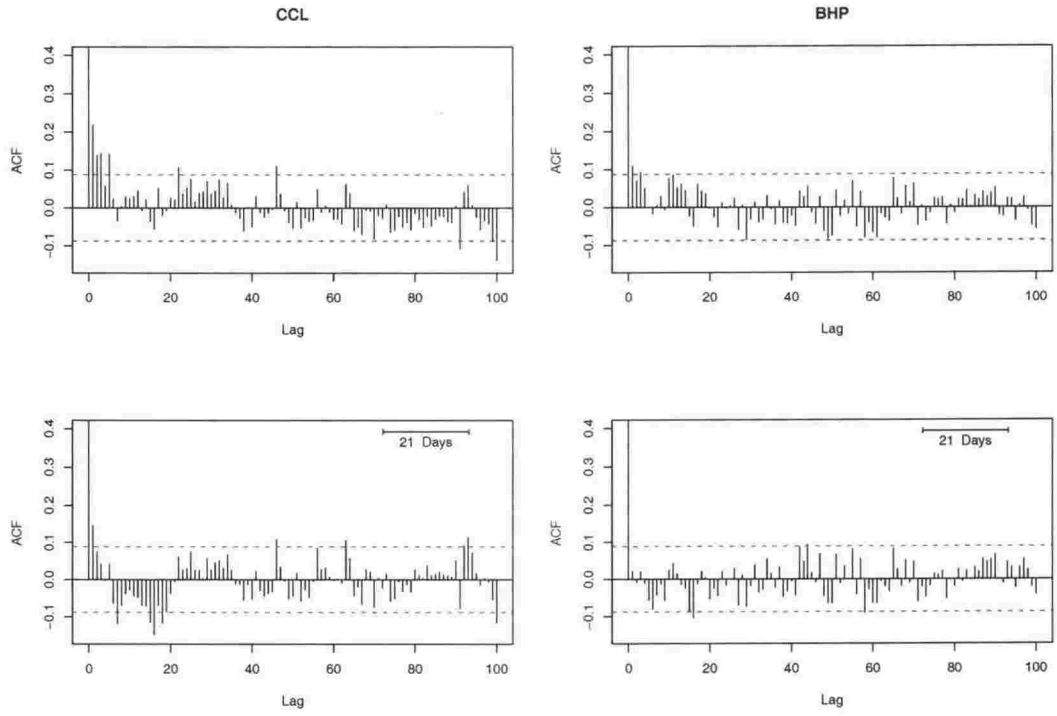

Figure 3.14. The estimated autocorrelation functions of absolute returns, and absolute standardised returns, for CCL on the left, and BHP on the right. The top plots are for $|R_t|$, and the lower for $|R_t - \hat{\mu}_t| / \hat{\sigma}_t$ where $\hat{\mu}_t$ is given by `loess`, and $\hat{\sigma}_t$ is the volatility given by the iterated t-volatility estimator. Approximate 95% confidence intervals for the autocorrelation estimates are shown. The four plots each have the same scale, and only $\rho_0 = 1$ is not shown.

where significant negative autocorrelation has been induced at low lags. There has been very little change in the correlogram of the BHP returns due to standardisation. The unimpressive changes due to standardisation for these two series do not affect our conclusions about the quality of the volatility estimates.

3.5 Conclusions

In this chapter, we have addressed non-parametric estimation of evolving volatility in the context of heavy-tailed distributions of returns. A new robust time series estimation procedure based on finite moving averages and the t-distribution has also been introduced. Motivated by the results of Chapter 2, the biweight A-estimator with auxiliary scale estimator Q_n and scaling constant $c = 11$ has been used to obtain robust volatility estimates that will be highly efficient for a range of distributions. In particular, simulation of daily returns, with a continuous volatility function, indicates that local volatility estimation based on this biweight A-estimator provides reliable estimates for the range of distributions encountered in empirical studies on financial returns: the t-distributions with ν close to 5, and also for normally distributed returns. By optimising volatility estimation for this target distribution (t_5), and benchmarking it against the all-purpose A-estimator, we have obtained an estimator which performed best in the many cases where the distribution of returns is well-approximated by a t-distribution.

Application of this iterated t-volatility estimator to real data provides sensible volatility estimates that are not unduly affected by occasional outlying returns. In cases where the returns have heavy tails, these estimates are shown to be consistent with the underlying distribution of the standardised returns. When the standardised returns are approximately normal, the volatility estimates do not differ greatly from traditional historical volatility estimates.

We conclude that the iterated t-volatility estimator, with $\nu = 5$, is a reliable estimator of volatility for daily financial price data. We feel confident that it will not only prevent extreme returns from having an undue influence on the volatility estimates, but also provide reliable estimates when the data is well behaved. Having secured a volatility estimation procedure, in the following chapter, we examine the empirical relationships between volatility and price level, commonly referred to as leverage effects.

Chapter 4

Leverage effects and a model for stock price

In this chapter, we investigate parametric option pricing models, and in particular the volatility functions they assume. We view the analysis undertaken at the end of this chapter as exploratory; if the relationships we (non-parametrically) identify are consistent with the stock price process assumed by a particular option pricing model, then that option pricing model may be more appropriate for options on that particular stock than competing models. It is for this reason that we restrict attention to stock price processes with known, closed form option pricing formulae.

All option pricing models assume a continuous time stock price process with volatility defined as follows.

Definition 4.1 (Stock price volatility) *We assume that stock price process S_t is a continuous time process with stochastic differential equation*

$$\frac{dS_t}{S_t} = \mu(S_t, t)dt + \sigma(S_t, t)dW_t$$

where $\mu(S_t, t)$ is the continuously compounding mean return, W_t is a Brownian motion process, and $\sigma(S_t, t)$ is defined to be the volatility of the stock at t.

When a firm has debt, the presence of this debt is likely to influence the volatility of the stock price; a phenomenon known as a leverage effect.

Leverage effects were first documented by Black (1976) with the empirical observation that as share price increases, volatility tends to decrease, and when share price decreases, volatility tends to increase. Black offers two explanations for this behaviour. The first is of "financial leverage" , i.e., when debt obligations are constant

regardless of variation in equity value. When the value of the firm falls, unlevered cash flows tend to fall. Interest payments are fixed, so, a given dollar fluctuation in unlevered cash flow exerts a greater percentage effect on cash flow net of interest payments, and hence a greater effect on equity value. Thus, equity volatility is greater. With constant interest payments, the reduction in equity value induces an increase in the leverage ratio, defined below.

Definition 4.2 (Leverage Ratio) *The leverage ratio of a firm is*

$$LR_t = \frac{B_t}{S_t} \tag{4.1}$$

where B_t is the value of the debt of the firm at t and S_t the value of the equity of the firm at t.

The opposite effect occurs when equity value rises, i.e, leverage rises and volatility falls. Thus, financial leverage results in an inverse relationship between stock price and stock price volatility.

Christie (1982) offers a formalisation of Black's (1976) leverage effect. Assuming the simple decomposition for firm value

$$V_t = S_t + B_t,$$

the instantaneous rate of return on the firm's (homogeneous) assets is

$$\frac{dV_t}{V_t} = \frac{S_t}{V_t}\frac{dS_t}{S_t} + \frac{B_t}{V_t}\frac{dB_t}{B_t} \tag{4.2}$$

where $\frac{dS_t}{S_t}$ and $\frac{dB_t}{B_t}$ are the instantaneous rates of return on the firm's equity and (risk-free) debt respectively.

Theorem 4.1 *Consider a firm with risk-free debt, earning a deterministic, continuously compounding rate of return, with value B_t at t, and homogeneous assets whose value has constant volatility σ. The stock price volatility for this firm is given by*

$$\sigma(S_t, t) = \sigma(1 + LR_t) \tag{4.3}$$

where LR_t is the leverage ratio for the firm, defined in (4.1).

Proof From (4.2), we see that

$$\frac{dS_t}{S_t} = \left(1 + \frac{B_t}{S_t}\right)\frac{dV_t}{V_t} - \frac{B_t}{S_t}\frac{dB_t}{B_t}$$

since $V_t = S_t + B_t$. Conditioning on time t values, noting the definition of LR_t and taking the standard deviation of both sides, we find

$$\sigma(S_t, t) = \sigma(1 + LR_t)$$

since σ is the constant volatility of the value of the firm, and the rate of return on risk-free debt is not stochastic. □

Since the leverage ratio is non-negative, and we assume that debt is risk-free and fixed (so that $V_t \geq B_t$), as $S_t \to \infty$ and the leverage ratio becomes very small, stock price volatility converges to a finite lower bound σ. However as $S_t \to 0$ and the leverage ratio becomes infinite, stock price volatility becomes infinite. Further, we note that volatility is a monotonic increasing function of financial leverage LR_t which in turn is inversely proportional to stock price, and so there is a negative relationship between stock price and stock price volatility.

This negative relationship can be characterised by the elasticity of the stock price volatility with respect to the stock price level.

Definition 4.3 (Elasticity of volatility) *The elasticity of stock price volatility, understood to be with respect to stock price level, is*

$$\theta_S = \frac{\partial \ln \sigma(S_t, t)}{\partial \ln S_t}$$

where the partial derivative is taken with respect to $\ln S_t$*, holding fixed all other arguments and parameters of the stock price volatility* $\sigma(S_t, t)$*.*

The elasticity of volatility for the simple leverage model is

$$\theta_S = \frac{-LR_t}{1 + LR_t} = -\frac{B_t}{V_t}. \tag{4.4}$$

Since $0 \leq B_t \leq V_t$ and B_t is assumed fixed, it follows that $-1 \leq \theta_S \leq 0$, and that as $S_t \to \infty$, $\theta_S \to 0$ and as $S_t \to 0$, $\theta_S \to -1$. The elasticity describes approximately the percentage change in stock price volatility for a 1% change in stock price level. To see this, we apply the chain rule of differentiation, and write

$$\frac{\partial \ln \sigma(S_t, t)}{\partial \ln S_t} = \frac{\partial \sigma(S_t, t)}{\partial S_t}\frac{S_t}{\sigma(S_t, t)} = \frac{\partial \sigma(S_t, t)}{\sigma(S_t, t)} \Big/ \frac{\partial S_t}{S_t}$$

and interpret the numerator and denominator as percentage changes in volatility and price respectively.

Returning to (4.4), we see that when stock price level is high, a 1% change in this level results in a very small percentage change in stock price volatility, however when S_t is small, in the presence of risk-free debt, there is a larger impact on volatility. The elasticity, and hence the size of the effect, is monotonic in S_t (and V_t), however the relationship is non-linear.

Black's (1976) second explanation for the leverage effect is "operating leverage", where the fixed costs of the firm have a similar effect to debt in the financial leverage story.

Throughout this chapter, we will refer to *any* relationship between stock price volatility and price level as a leverage effect, whether or not it is modeled using debt. When the effect is directly consistent with the effects described by Black (1976), Christie (1982) and others, it will be referred to as the classical leverage effect.

4.1 Parametric modelling of the leverage effect

In this section we describe various attempts to (implicitly or explicitly) model leverage effects, and we discuss their consistency with the classical leverage effect.

4.1.1 The constant elasticity of variance option pricing model

The first attempt to incorporate leverage effects into an option pricing model came from the constant elasticity of variance (CEV) model (Cox & Ross 1976), in which a mathematical relationship is proposed for volatility in terms of stock price. Cox (1996) describes the motivation for the model being a request from Black (1976) to model the empirical relationships he observed in stock price data. The CEV solution can be regarded as a statistical model for the underlying stock price process, rather than a financial one. It uses a mathematical specification which reflects the classical leverage behaviour, without actually acknowledging debt.

The underlying CEV stock price process (hereafter referred to as the CEV process) is specified by the SDE

$$\frac{dS_t}{S_t} = \mu dt + \delta S_t^{\frac{\beta-2}{2}} dW_t \tag{4.5}$$

and it gets its name from the fact that the elasticity of stock price variance (volatility squared) with respect to stock price level is constant. To see this, we first note that

$$\sigma(S_t, t) = \delta S_t^{\frac{\beta-2}{2}} \tag{4.6}$$

is the volatility of the CEV process, and calculate the elasticity of volatility

$$\theta_S \equiv \frac{\partial \ln \sigma(S_t, t)}{\partial \ln S_t} = \frac{1}{2}(\beta - 2).$$

The elasticity of variance is just two times the elasticity of volatility, and this is a constant. The CEV process was originally specified for $0 \leq \beta < 2$, consistent with the notion of financial leverage in (4.3). The upper limiting case of $\beta = 2$ (corresponding to the Black-Scholes model) was excluded from the model due to the different mathematical behaviour of the solution in this case.

The CEV volatility function is shown for $\beta = 0$, 1 and the limiting case of $\beta = 2$ in Figure 4.1. These three cases cover the range of β whose elasticities are consistent with the classical leverage effect, and include the null case of constant volatility (GBM, with $\beta = 2$). The explosive nature of volatility as S_t nears 0 for the two processes with $\beta < 2$ is evident, as well as the inverse relationship between stock price and volatility. The cases all have the same volatility at $S_t = 1$, and for fixed S_t, volatility is monotonically increasing in β for $S_t < 1$ and decreasing for $S_t > 1$.

Despite its motivation, the CEV model makes no explicit allowance for debt in the model. The model (4.5) facilitates a known transitional density for future price $S_{t+\tau}$ given S_t, and a closed form solution for call option price. The former is a mixed distribution consisting of a positive probability that bankruptcy has occurred by $t + \tau$, and a continuous distribution for $S_{t+\tau} > 0$. Although the continuous part of the density is not of a standard distribution, Schroder (1989) utilises the distribution function of the non-central chi-squared random variable. He writes the call option price, found by evaluation of

$$C_t = e^{-r\tau} \mathrm{E}_t^{\mathbb{Q}} \{\max(S_{t+\tau} - K, 0)\}$$

where C_t is the price of the option, K is the exercise price of the option, r is the continuously compounding risk-free rate, and S_t follows the CEV model (4.5) in the risk-neutral probability measure $\mathbb{Q}$ (see Harrison & Pliska (1981) for discussion) with $\mu = r$, in terms of non-central chi-squared probabilities. This reduces an option pricing formula featuring infinite sums to something akin to the Black-Scholes (1973)

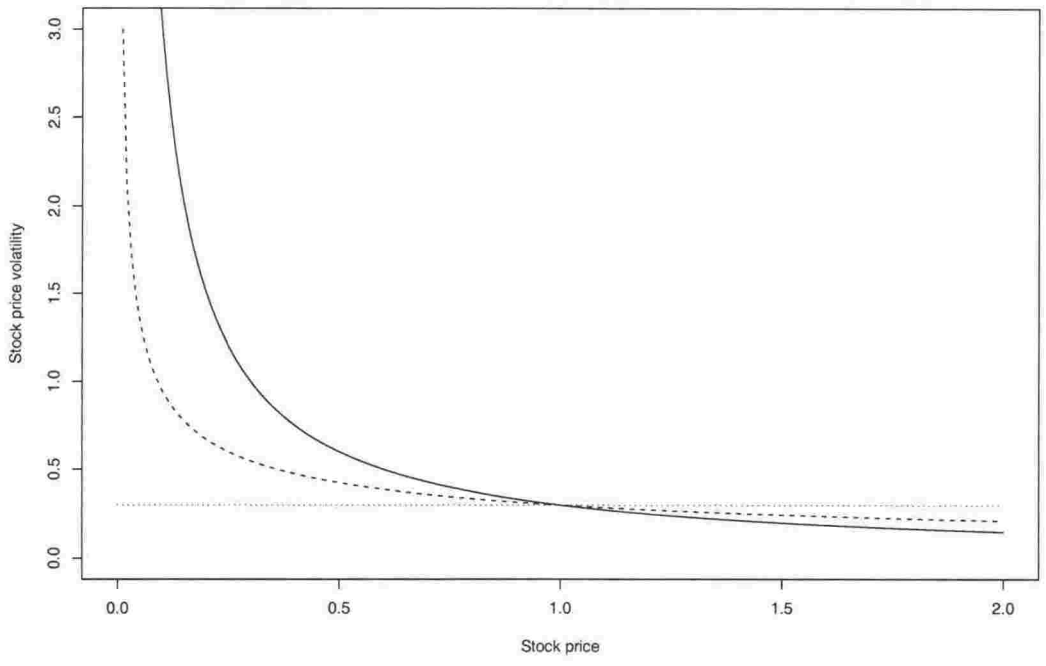

Figure 4.1. The constant elasticity of variance stock price volatility function, defined in (4.6). The solid line is the function for the absolute CEV process ($\beta = 0$), the dashed line for the square root CEV process ($\beta = 1$) and the dotted line for GBM ($\beta = 2$). All processes have $\delta = 0.3$.

formula. In addition, the mathematical solution for call price for the cases $\beta < 0$ and $\beta > 2$ are well defined (see Emanuel & MacBeth (1982) for the latter case or Randal (1998) for discussion).

The benefits of the CEV model are largely mathematical: GBM and the Black-Scholes formula are contained as a special case, and the extra parameter allows a leverage effect between volatility and stock price that includes the classical situation but is otherwise well defined for any $\beta \in \mathbb{R}$. While it has this flexibility, unlike the compound option pricing model discussed in the following section, even for $0 \leq \beta < 2$, the CEV model does not obey the elementary limiting behaviour of (4.3), since as $S_t \to \infty$, $\sigma(S_t, t) \to 0$ rather than to a positive constant.

4.1.2 The compound option pricing model

The stock price volatility for the stock price process implied by the compound option pricing model (Geske 1979) is consistent with the limiting properties of volatility for the risk-free debt model that the CEV model fails in regard to. Geske assumes that firm value follows GBM, with

$$\frac{dV_t}{V_t} = \mu dt + \sigma dW_t$$

and hence has constant volatility. In the case where the firm has zero debt, $S_t = V_t$ and so, like the CEV model, the Black-Scholes formula is a special case of the option pricing formula.

This compound model formally acknowledges risky debt, in the form of a single fixed payment M made at time $t_d > t$. At time t_d, the creditors are paid using the realised value of the firm, however limited liability implies that the actual payment is $\min(V_{t_d}, M)$. Consequently, at t_d, the equity of the firm is worth

$$S_{t_d} = \begin{cases} V_{t_d} - M & V_{t_d} > M \\ 0 & V_{t_d} \leq M. \end{cases}$$

This model recognises that the creditors may not get all their money back, and is thus a more realistic model than the risk-free debt model.

Geske outlines an argument which shows that the value of the stock at time t is given by the Black-Scholes formula. This follows the derivation of the Black-Scholes equation using Itô's Lemma, the construction of a hedge portfolio, and the boundary condition above. In particular, the stock price at any time $t < t_d$ is given by the Black-Scholes formula with firm value V_t as the underlying asset, and M as the exercise price:

$$S_t \equiv S(V_t, t) = V_t \Phi(g_t) - M e^{-r\tau_d} \Phi(g_t - \sigma\sqrt{\tau_d}) \tag{4.7}$$

where

$$g_t = \frac{\ln V_t - \ln M + (r + \frac{1}{2}\sigma^2)\tau_d}{\sigma\sqrt{\tau_d}}, \tag{4.8}$$

$\tau_d = t_d - t$ is the time-to-maturity of the debt, σ is the constant volatility of the firm's assets and $\Phi(x)$ is the standard normal cumulative distribution function.

Since the firm's assets follow GBM, $dV_t = \mu V_t dt + \sigma V_t dW_t$ and applying Itô's Lemma to the function $S(V_t, t)$ given in (4.7), we obtain

$$dS_t = \frac{\partial S_t}{\partial V_t} dV_t + \frac{\partial S_t}{\partial t} dt + \frac{1}{2}\sigma^2 V_t^2 \frac{\partial^2 S_t}{\partial V_t^2} dt.$$

In particular, the volatility of the stock is given by the coefficient of $S_t dW_t$ in the partial differential equation above, and is

$$\sigma(S_t, t) \equiv \sigma_S(V_t, t) = \sigma \frac{V_t}{S_t} \frac{\partial S_t}{\partial V_t} = \sigma \frac{V_t}{S(V_t, t)} \Phi\left(\frac{\ln V_t - \ln M + (r + \frac{1}{2}\sigma^2)\tau_d}{\sigma\sqrt{\tau_d}}\right) \tag{4.9}$$

where we use the familiar Black-Scholes hedge ratio

$$\frac{\partial S_t}{\partial V_t} = \Phi(g_t) \tag{4.10}$$

and acknowledge that the underlying variable in this case is V_t not S_t.

Unlike for the CEV model, $\sigma(S_t, t)$ is a function of time through the time-to-maturity of the debt. As explained by Geske,

$$\frac{\partial S_t}{\partial \tau_d} = Me^{-r\tau_d}\left[\frac{\sigma\phi(g_t - \sigma\sqrt{\tau_d})}{2\sqrt{\tau_d}} + r\Phi(g_t - \sigma\sqrt{\tau_d})\right] > 0 \tag{4.11}$$

and so as time-to-maturity of the debt decreases, given no change in V_t, so too does S_t. This decrease in S_t will serve to increase financial leverage and thereby increase volatility. Because time is monotonic, this effect will always operate, regardless of the general movement in firm value. A decrease in V_t will amplify the increase in volatility; however, when firm value increases, the resulting decrease in volatility may be offset.

Although the form of (4.9) is not directly consistent with (4.3) due to the probability $\Phi(g_t)$, its limiting behaviour is, as described in the following theorem.

Theorem 4.2 *The stock price volatility under the compound option pricing model, $\sigma_S(V_t, t)$, defined in (4.9), has the following properties:*

1. $\sigma_S(V_t, t) \geq \sigma$;
2. *As* $V_t \to \infty$, $\sigma_S(V_t, t) \to \sigma$;
3. *As* $V_t \to 0$, $\sigma_S(V_t, t) \to \infty$.

Theorem 4.2 shows that stock price volatility in the compound option pricing model is consistent with the basic behaviour implied for a firm with leveraged equity and risk-free debt, and its proof can be found in Appendix D.1. We also note that under both the compound model (with risky debt) and the risk-free debt model, the stock is more volatile than the firm, indicating that the presence of debt transfers risk from debtholders to the stockholders, whether the debt is risk-free or not, i.e. stockholders bear a greater proportion (than debtholders) of the risk associated with the firm's assets. Since $\frac{\partial S_t}{\partial V_t}$ is a probability, we note that for the compound model

$$\sigma \leq \sigma_S(V_t, t) < \sigma(1 + LR_t)$$

where $(1 + LR_t) = \frac{V_t}{S_t}$. This contrasts with (4.3), and we conclude that some of the risk due to debt is borne by the holders of the risky-debt.

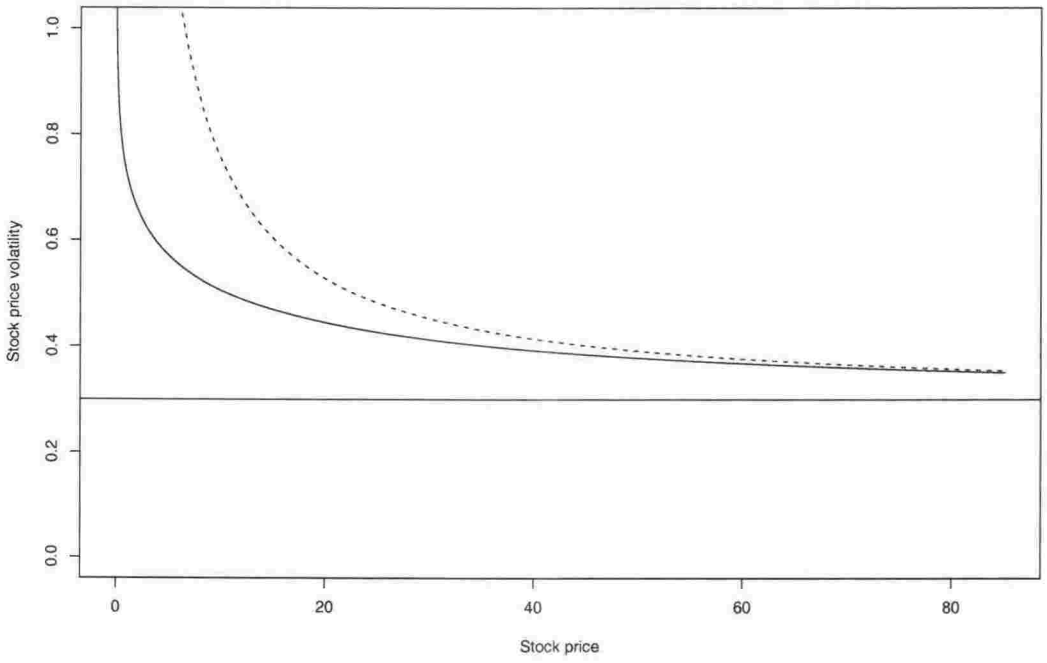

Figure 4.2. The volatility function for the compound option pricing model, defined in (4.9), plotted against S_t, defined in (4.7), with $M = 25$, $\tau_d = 10$, $r = 0.05$ and $\sigma = 0.3$. Also shown using the dotted line is the volatility when debt is risk-free, defined in (4.3), where $B_t = Me^{-r\tau_d}$.

An example of the volatility function for a firm with $M = 25$ due in $\tau_d = 10$, and with $r = 0.05$ and $\sigma = 0.3$, is shown in Figure 4.2 and compared to that under the risk-free debt model. The lower bound σ is shown, and the explosive nature of the volatility when S_t nears zero under both debt models is clear.

As before, we also consider the elasticity of volatility (with respect to stock price) for this model. Care must be taken, since the underlying variable in this case is firm value V_t. Since the stock price is a 1-1 function of firm value

$$\frac{\partial \ln \sigma_S(V_t,t)}{\partial \ln S(V_t,t)} = \frac{\partial \sigma_S(V_t,t)}{\partial S(V_t,t)} \frac{S(V_t,t)}{\sigma_S(V_t,t)} = \left[\frac{\partial \sigma_S}{\partial V_t} \Big/ \frac{\partial S_t}{\partial V_t}\right] \frac{S(V_t,t)}{\sigma_S(V_t,t)}$$

and following from (4.9),

$$\frac{\partial \sigma_S}{\partial V_t} = \sigma \left[\frac{1}{S_t}\frac{\partial S_t}{\partial V_t} - \frac{V_t}{S_t^2}\left(\frac{\partial S_t}{\partial V_t}\right)^2 + \frac{V_t}{S_t}\frac{\partial^2 S_t}{\partial {V_t}^2}\right]. \tag{4.12}$$

This equation is not directly consistent with the function for $\frac{\partial \sigma_S}{\partial S}$ given by Geske, who ignores the functional relationship between S_t and V_t. In particular, if we take (4.9) and allow S_t to vary while fixing V_t, then differentiate with respect to S_t, we obtain

$$\frac{\partial \sigma_S}{\partial S_t} = -\sigma \frac{V_t}{S_t^2}\frac{\partial S_t}{\partial V_t}$$

as given by Geske (Geske 1979, page 73). This corresponds to the second term in (4.12), which when multiplied by the remaining terms to give the elasticity, is consistent with a CEV process.

Following from (4.12), the correct elasticity of volatility with respect to stock price for the compound model is

$$\theta_S(V_t,t) \equiv \left[\frac{\partial \sigma_S}{\partial V_t} \Big/ \frac{\partial S_t}{\partial V_t}\right] \frac{S(V_t,t)}{\sigma_S(V_t,t)} = \frac{S(V_t,t)}{V_t \frac{\partial S_t}{\partial V_t}} + S(V_t,t) \frac{\frac{\partial^2 S_t}{\partial V_t^2}}{(\frac{\partial S_t}{\partial V_t})^2} - 1. \tag{4.13}$$

where $\sigma_S(V_t,t)$ is substituted from (4.9). Evaluating partial derivatives, substituting in (4.13) and simplifying, we find

$$\theta_S(V_t,t) = \frac{-Me^{-r\tau_d}\Phi(g_t - \sigma\sqrt{\tau_d})}{V_t\Phi(g_t)} + \frac{S_t\phi(g_t)}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)^2}. \tag{4.14}$$

where S_t is given in (4.7), g_t is defined in (4.8), and $\phi(x)$ is the standard normal probability density function. The following theorem describes some of the properties of this elasticity.

Theorem 4.3 *The elasticity of stock price volatility under the compound option pricing model, $\theta_S(V_t,t)$, given in (4.14), has the following properties:*

1. $\theta_S(V_t,t) > \max[-1, -(Me^{-r\tau_d})/V_t]$;
2. *As* $V_t \to \infty$, $\theta_S(V_t,t) \to 0$;
3. *As* $V_t \to 0$, $\theta_S(V_t,t) \to 0$.

A proof to this theorem is given in Appendix D.1. A further result, for which we cannot provide a proof, is given in the following conjecture.

Conjecture 4.4 *The elasticity of stock price volatility under the compound option pricing model, $\theta_S(V_t,t)$, defined in (4.14), is always negative.*

Theorem 4.3 and Conjecture 4.4 summarise the properties of the elasticity of stock price volatility under the compound option pricing model. Comparison with the results for the simple risk-free debt model shows similar behaviour as firm value (and stock value) increases; however, as firm price approaches zero (or indeed $Me^{-r\tau_d}$) elasticity declines in absolute value, with a limit of zero. This behaviour reflects the

transfer of risk from the stockholders (who bear almost all risk when firm value is high) to the debtholders. Conjecture 4.4 implies that the compound option model is always consistent with the classical leverage effect.

The behaviour described in Theorem 4.3 and Conjecture 4.4 is evident in Figure 4.3 for three different levels of debt. In particular, for the chosen parameters, the limit as $V_t \to 0$ is clear, and for the smaller values of M, the limit as $V_t \to \infty$ is also apparent. When $M = 50$, the lower bound is shown, and this becomes a very good approximation to the elasticity when V_t is much larger than M. In addition to the stated properties, we also note that as V_t increases from zero, the elasticity function is concave down, i.e., it decreases at an increasing rate. This reflects an increasingly large decrease in volatility for a fixed percentage change in V_t. At a point near or at $V_t = Me^{-r\tau_d}$, the second derivative of elasticity changes sign, and the function becomes concave upward for large values of V_t. This indicates the rate of change of elasticity is slowing, which means the change in volatility for a fixed percentage change in V_t is still growing, but at a decreasing rate. As V_t grows beyond M, this rate of change reaches a minimum, and then starts to converge to zero. The behaviour of the first and second derivatives of elasticity reflect the transfer of risk between debt-holders and stock-holders, and we see that when V_t is large enough, all risk is borne by the stock-holders, as in the risk-free debt model.

Since $\theta_S(V_t, t)$ is also a function of τ_d, we can also examine its properties as a function of t. An intuitively appealing result is given in the following theorem.

Theorem 4.5 *The elasticity of stock price volatility under the compound option pricing model, $\theta_S(V_t, t)$, defined in (4.14), has the following limit*

$$\lim_{\tau_d \to 0} \theta_S(V_t, t) = \begin{cases} 0 & V_t < M \\ -\frac{M}{V_t} & V_t > M. \end{cases}$$

The proof to Theorem 4.5 can be found in Appendix D.1. The appeal of this result comes through comparison with the case of risk-free debt. As $\tau_d \to 0$, if $V_t > M$ the debt is, for all practical purposes, risk-free, and the elasticity given in the theorem matches (4.4). When $V_t < M$, and $\tau_d \to 0$, stock holders have zero claim on the firm, and their holdings have zero volatility. Thus the elasticity is also zero as given in Theorem 4.5. This behaviour is demonstrated graphically in Figure 4.4. As $\tau_d \to 0$, we see the elasticity function converging to the correct limits. The elasticity

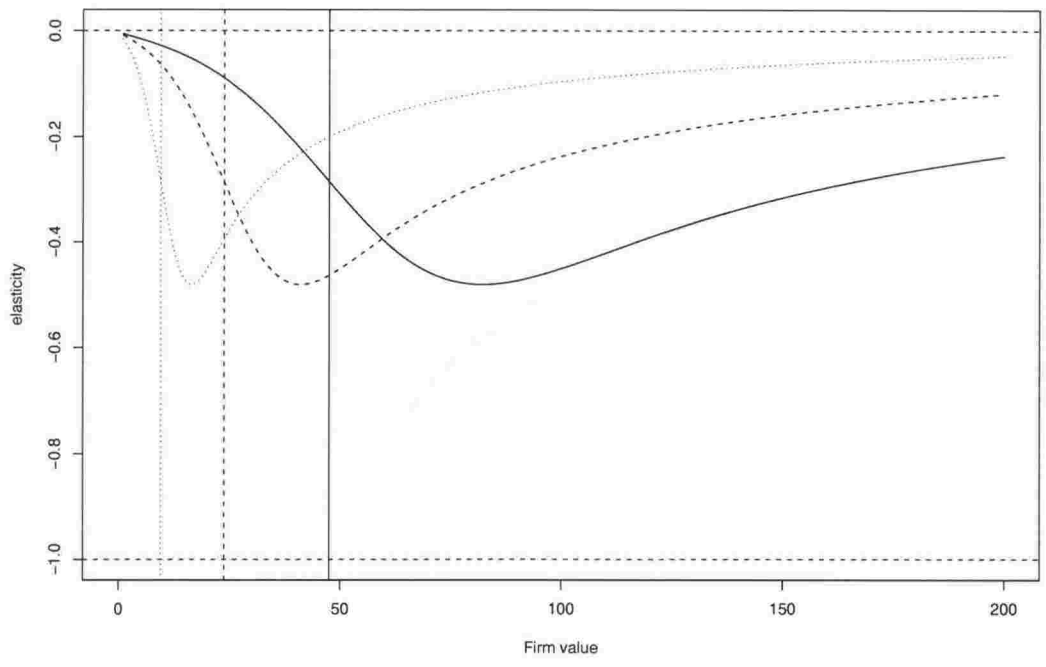

Figure 4.3. The elasticity of stock price volatility for the compound option pricing model as a function of V_t and for three different choices of M, with $\tau_d = 1$, $r = 0.05$ and $\sigma = 0.3$. The solid line is for $M = 50$, the dashed for $M = 25$ and the dotted line for $M = 10$. The limits of 0 and -1 are shown by the dotted lines, and the lower bound for $M = 50$ only, is shown in grey. The vertical lines are at $Me^{-r\tau_d}$ for the three curves.

is plotted again in Figure 4.5, however this time against S_t. The left hand part of the function in Figure 4.4 is condensed, since $S_t \approx 0$ for all values of $V_t < M$. As a function of S_t, elasticity converges to (4.4) as required.

Thus, in conclusion, the volatility process for the compound option pricing model exhibits behaviour broadly consistent with the risk-free debt model, and Black's (1976) empirical observations. The elasticity of this volatility with respect to stock price is negative, but not a constant function of S_t as it was in the CEV model. The actual shape of the elasticity function reflects risk transfer between the stockholders and debtholders of the firm.

4.1.3 The displaced diffusion model

The displaced diffusion option pricing model (Rubinstein 1983) is described as a leverage model (see for example Bates 2000) because it models the stock in the presence of debt; however the behaviour of volatility in this model is not consistent with the classical leverage effect documented by Black (1976). Rather, as stock price increases, volatility also increases.

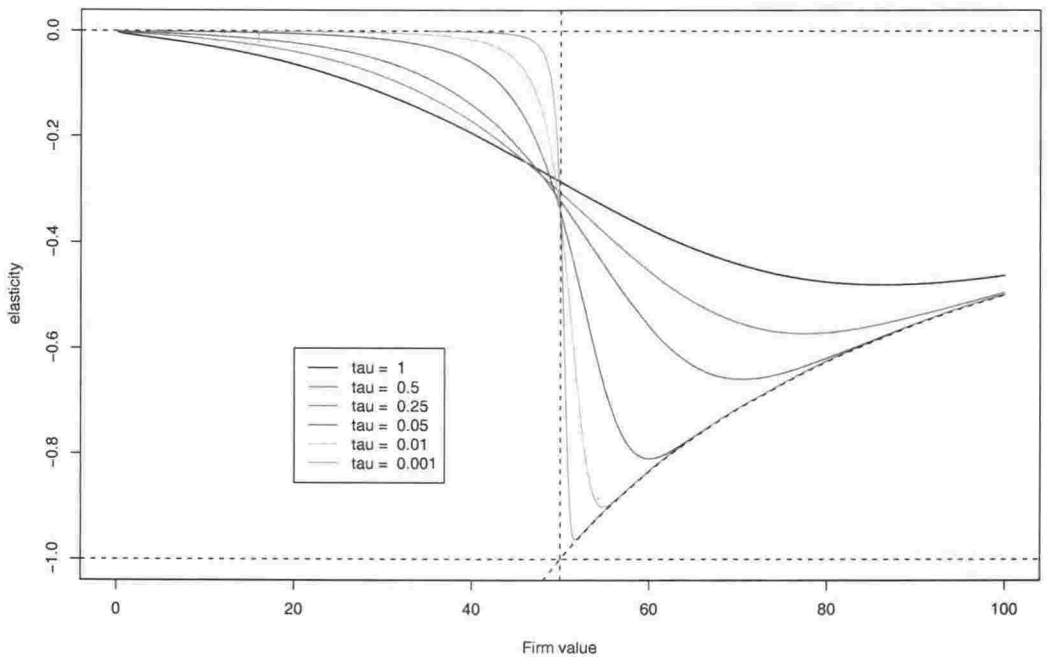

Figure 4.4. The elasticity of stock price volatility for the compound option pricing model as a function of V_t and τ_d. In all cases, $M = 50$, $r = 0.05$ and $\sigma = 0.3$, with the function plotted for each τ_d in $\{1, 0.5, 0.25, 0.05, , 0.01, 0.001\}$ as described in the legend. The limits of 0 and -1 are shown by the dotted lines, as is the lower bound for $V_t > M$. The vertical dotted line is drawn at M.

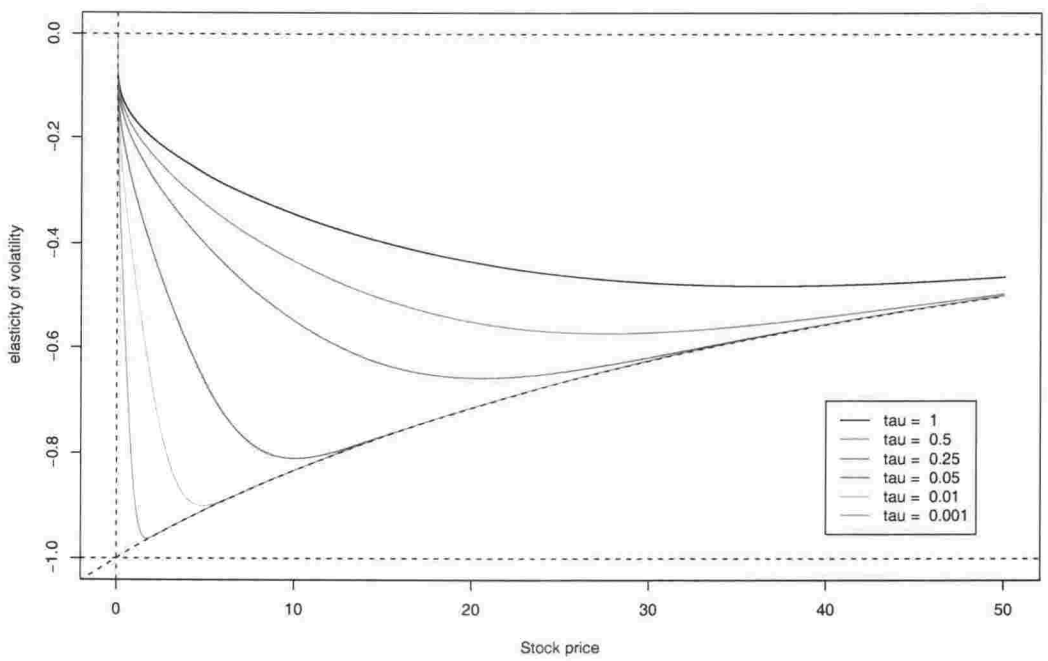

Figure 4.5. The elasticity of stock price volatility for the compound option pricing model as a function of S_t and τ_d. In all cases, $M = 50$, $r = 0.05$ and $\sigma = 0.3$, with the function plotted for each τ_d in $\{1, 0.5, 0.25, 0.05, , 0.01, 0.001\}$ as described in the legend, and these are used to calculate both $\theta_S(V_t, t)$ and $S(V_t, t)$ for a range of V_t. The limits of 0 and -1 are shown by the dotted lines, as is the lower bound for $S_t > 0$.

The displaced diffusion option pricing model extends the firm (from the Black-Scholes assumptions) to allow two sorts of assets: risky assets whose value follows GBM, and non-risky assets which compound at the risk-free rate. We assume that at time $t = 0$, the initial value of the firm V_0 is invested with proportion α_0 into risky assets whose value evolves according to GBM, and the remaining proportion into assets with no associated risk. The risk-free assets compound continuously at the risk-free rate r, and hence their value at time t is $R_t = R_0 e^{rt}$ where $R_0 = (1 - \alpha_0)V_0$ is the initial investment in the risk-free assets.

Let A_t be the value at time t of the risky portion of the firm. This is assumed to follow GBM, and is given by the standard solution, i.e., under the risk-neutral measure $\mathbb{Q}$, risky asset value at t is

$$A_t = A_0 \exp\{(r - \tfrac{1}{2}\sigma^2)t + \sigma W_t\}$$

for all t, where $A_0 = V_0 - R_0 = \alpha_0 V_0$ is the initial investment in the risky assets, r is the continuously compounding risk-free rate, σ is the (constant) volatility of the risky assets, and W_t is a Brownian motion process under $\mathbb{Q}$. At time t, the firm has value

$$V_t = A_t + R_t \tag{4.15}$$

and since A_t is a GBM process, $A_t > 0$ for all t, and hence $V_t > R_t$.

We define $\alpha_t \equiv \frac{A_t}{V_t}$ to be the proportion of the value of the firm invested in the risky assets at time t, and note that this is a stochastic process with

$$\alpha_t = 1 - \frac{1}{1 + \frac{\alpha_0}{1-\alpha_0} \exp\{-\frac{1}{2}\sigma^2 t + \sigma W_t\}}$$

where W_t is a Brownian motion process. Although stochastic, due to the nature of R_t, α_t satisfies

$$(1 - \alpha_t)V_t = (1 - \alpha_0)V_0 e^{rt} \tag{4.16}$$

for all t. As noted by Rubinstein, given the path of V_t, this property allows us to determine α_t only once in the firm's history in order to obtain the path of α_t for all t.

In addition to the assumption of heterogeneous assets, risk-free debt worth B_t at t is allowed. In order to ensure the debt is risk-free, a restriction is imposed on the level of debt relative to the risk-free assets of the firm. The assumption is that $B_t \leq R_t$, and this ensures bankruptcy is impossible. Since the debt is risk-free, the value of

this debt at time t is given by $B_t = B_0 e^{rt}$ for all t. The firm's equity value at time t is given by

$$S_t = V_t - B_t$$

where V_t is defined in (4.15).

Given the simple decomposition of firm value into stock and risk-free debt, we might expect stock price volatility to be exactly consistent with the classical leverage effect. The actual behaviour depends on our underlying variable, and is due to the presence of risk-free assets. If we consider total firm value (risky + non-risky assets) to be the underlying variable, and vary V_t while keeping α_t fixed, we force change in R_t, and the behaviour of volatility is consistent with (4.3). However, this situation is not sensible due to the non-risky assets. Rather, we must fix R_t, and vary V_t through A_t alone. As a consequence α_t varies, and we see quite different behaviour in the volatility.

Since A_t is a GBM process under the risk-neutral measure $\mathbb{Q}$, we know

$$dA_t = rA_t dt + \sigma A_t dW_t$$

where W_t is a Brownian motion process under $\mathbb{Q}$. Firm value V_t is given by (4.15), and so we have

$$dV_t = dA_t + rR_t dt = rV_t dt + \sigma(V_t - R_t)dW_t$$

since $A_t = V_t - R_t$. By the identity $S_t = V_t - B_t$,

$$dS_t = dV_t - rB_t dt = rS_t dt + \sigma(V_t - R_t)dW_t$$

and by the definition of stock price volatility, we note

$$\sigma_S(V_t, t) = \sigma \frac{V_t - R_t}{S_t} \tag{4.17}$$

where $V_t - R_t = A_t$ follows a GBM process, and $S_t = V_t - B_t$. Treating R_t and B_t as fixed, we write

$$\sigma_S(V_t, t) = \sigma \left(1 - \frac{\Delta_t}{S(V_t, t)}\right) \tag{4.18}$$

where $\Delta_t = R_t - B_t = (R_0 - B_0)e^{rt}$ denotes the value of the risk-free assets in excess of the debt, with $\Delta_t \geq 0$ due to the assumption that debt is risk-free. We note that this function is *increasing* in S_t, and in particular, if $S_t \to \infty$, $\sigma_S(V_t, t) \to \sigma$ from below.

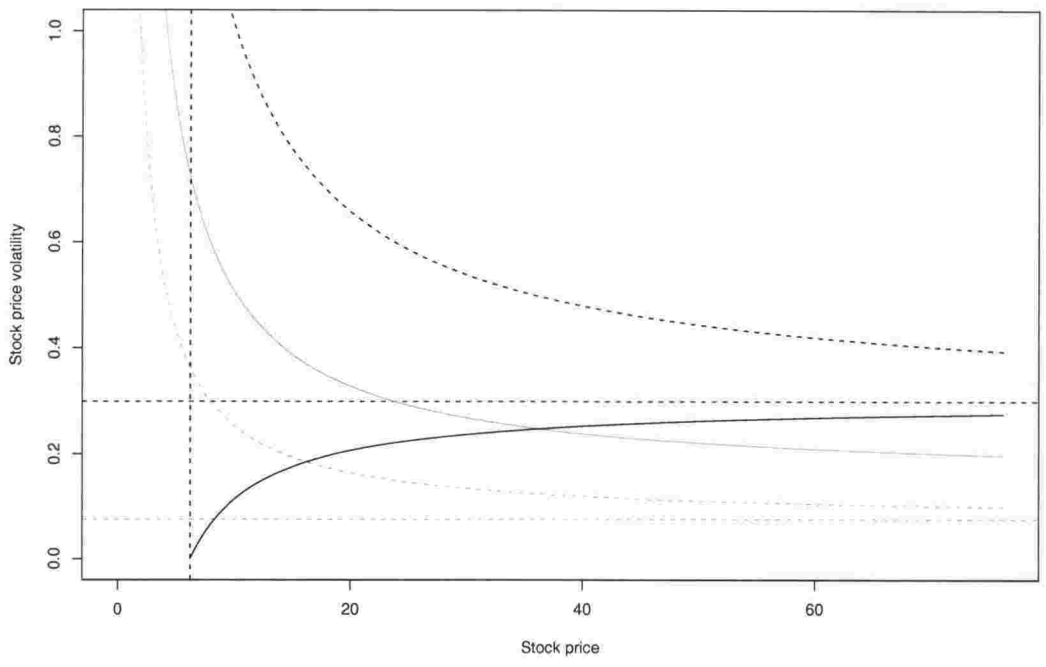

Figure 4.6. The volatility function for the displaced diffusion model, defined in (4.18), plotted (using the solid line) against S_t, with $B_t = 25$, $R_t = 30$, $r = 0.05$ and $\sigma = 0.3$. Also shown using the dashed line is the horizontal asymptote at $\sigma = 0.3$ of this function. The vertical dotted line marks $\Delta_t = R_t - B_t$. The grey solid line is (4.17) for $\alpha_t = 0.5$, and the grey dotted line is (4.17) for $\alpha_t = 0.25$. These functions have upper bound given by the curved dotted line, representing (4.3) and $\alpha_t = 1$. The asymptote for the curve with $\alpha_t = 0.25$ is shown by the grey dotted line.

Stock price volatility for the displaced diffusion model and selected parameters is shown in Figure 4.6. This figure shows the volatility function (4.18) for the applicable range $S_t \geq \Delta_t$, and this is given by the solid line. Since the displaced diffusion model is a risk-free debt model, the function (4.17) is also shown for fixed $\alpha_t \in \{1, 0.5, 0.25\}$. Although it is useful to reconcile the volatility for this risk-free debt model with that given in Theorem 4.1 for the general risk-free debt model, (4.17) implicitly ignores the presence of the risk-free assets. In particular, since S_t is a deterministic function of $A_t = \alpha_t V_t$ for fixed t, we cannot simultaneously fix α_t and vary S_t in this model. At $S_t = \Delta_t$, $V_t = R_t$ and both $A_t = 0$ and $\sigma_S(V_t, t) = 0$. As S_t increases (through increase in A_t), both volatility and α_t increase, and we see the volatility curve intersecting first the function (4.17) for $\alpha_t = 0.25$ and then for $\alpha_t = 0.5$.

The behaviour shown in (4.18) and Figure 4.6, implies a positive elasticity of volatility for this model. Since Δ_t is fixed, from (4.18), we can derive the elasticity of volatility

$$\theta_S \equiv \frac{\partial \ln \sigma_S(V_t, t)}{\partial \ln S(V_t, t)} = \frac{\sigma \Delta_t}{S_t^2} \frac{S_t}{\sigma_S} = \frac{\Delta_t}{S_t - \Delta_t} \tag{4.19}$$

where $S_t \geq \Delta_t$, and clearly, $\theta_S > 0$. As $S_t \to \Delta_t$, the elasticity becomes infinite, reflecting the fact that close to $S_t = \Delta_t$, a small percentage change in S_t (from Δ_t to

$\Delta_t + \epsilon$) results in a very large percentage increase in risk, i.e. from zero risk to some risk. For large S_t, the effect of the risk-free assets is negligible, and the change in volatility for a unit percentage change in S_t is close to zero, reflecting approximately constant volatility as $S_t \to \infty$.

4.1.4 Leverage models and the volatility smile

A popular method of estimating volatility is to imply it from stock and call option prices using the Black & Scholes (1973) option pricing model. This technique and the relevant literature are discussed in detail by Mayhew (1995). In its common usage "implied volatility" is the term used to describe the value for σ which can be used to equate the Black-Scholes formula to the observed market price. The Black-Scholes formula is

$$C_t = S_t \Phi(h_t) - K e^{-r\tau} \Phi(h_t - \sigma\sqrt{\tau}) \tag{4.20}$$

with

$$h_t = \frac{\ln S_t - \ln K + (r + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}$$

and where S_t is the stock price at time t, K is the exercise price of the option with time to maturity τ, r is the continuously compounding risk-free rate and σ is the constant volatility of the stock. The assumption of constant volatility can be relaxed to allow a deterministic function of time σ_t, with σ in (4.20) replaced by

$$\sigma = \left[\frac{1}{\tau}\int_t^{t+\tau} \sigma_u^2 du\right]^{\frac{1}{2}}$$

which is the square root of the average variance over the remaining life of the option.

The Black-Scholes equation (4.20) has five arguments; however, with the exception of σ, these are readily observable quantities. Given that market call option prices are also observable, and (4.20) is a monotonic function of σ, the Black-Scholes equation can be used to "imply" the parameter σ that would be used to produce the market price. The implied volatility has been described as "essentially a normalised option price" (Gourieroux & Jasiak 2001, page 323), whereby the strike price and time to maturity of the option are eliminated and a single summary statistic produced. Time series data of both S_t and C_t will yield a time series estimate of σ_t in the obvious way.

The primary theoretical assumption underpinning use of this statistic is that the stock price has constant (or deterministic) volatility. Provided this assumption is correct, the implied volatility reflects the option market's forecast of the average variance over the remaining life of the option, and impounds any information currently available in both the stock market and the option market. However, in order to benefit from information contained in the observed call price, we must be sure that market participants are indeed pricing options using the Black-Scholes formula, and that at any instant, observed prices differ from those given by (4.20) only because of differences in market participants' abilities to estimate an appropriate value of σ.

Issues raised in Mayhew's (1995) review of the literature on implied volatility should be enough to rule out use of the Black-Scholes implied volatility. In particular, implied volatility estimates themselves cast doubt on the most fundamental of the assumptions made by Black & Scholes: that the stock price follows GBM. Should the underlying asset price follow GBM, then any options with the same maturity on that underlying asset should provide the same implied volatility estimates. In practice this is not the case, and systematic patterns are obtained between implied volatility and both time to maturity, and strike price. In essence, empirical irregularities in the implied volatilities suggest the Black-Scholes model does not correctly price call options, which in turn implies that the underlying asset price does not have constant volatility. Despite the contradictions inherent in Black-Scholes implied volatilities, this method remains a compelling volatility estimation procedure. In order to make it operational, we should seek an option pricing model which gives estimates that are consistent with the underlying assumptions of that model.

One significant reason that the CEV, compound option, and displaced diffusion models have been described above is that all three facilitate closed form option pricing models, all similar in form to the Black-Scholes equation, and requiring numerical methods only to calculate probabilities as in Black-Scholes' $\Phi(h_t)$ and $\Phi(h_t - \sigma\sqrt{\tau})$. The probability functions required in the case of CEV and compound option models are certainly less common that the standard normal; however many modern statistical packages contain efficient algorithms for their approximation (just as the standard normal probabilities are numerically approximated in a much larger number of packages).

All option pricing models depend on the fundamental characteristics of the option: strike price K and time-to-maturity τ. In addition, because they are all based on

risk-neutral pricing, all depend on the continuously compounding risk-free rate r. Additional parameters depend on the underlying structural assumptions about the firm and stock price evolution. The Black-Scholes formula (4.20) has additional parameters S_t and $\sigma(S_t, t) = \sigma$.

The constant elasticity of variance option pricing formula (Cox & Ross 1976, Emanuel & MacBeth 1982, Schroder 1989) has additional parameters S_t, β and δ, as specified in the stochastic differential equation for this process (4.5), and gives

$$C_t = S_t P_1(S_t, K, \tau, r, \delta, \beta) - Ke^{-r\tau} P_2(S_t, K, \tau, r, \delta, \beta) \tag{4.21}$$

where

$$P_1(S_t, K, \tau, r, \delta, \beta) = \begin{cases} Q(2y; 2 + \frac{2}{2-\beta}, 2x) & \beta < 2 \\ \Phi(h_t) & \beta = 2 \\ Q(2x; \frac{2}{2-\beta}, 2y) & \beta > 2 \end{cases}$$

and

$$P_2(S_t, K, \tau, r, \delta, \beta) = \begin{cases} 1 - Q(2x; \frac{2}{2-\beta}, 2y) & \beta < 2 \\ \Phi(h_t - \sigma\sqrt{\tau}) & \beta = 2 \\ 1 - Q(2y; 2 + \frac{2}{2-\beta}, 2x) & \beta > 2 \end{cases}$$

where $Q(x; \nu, \lambda)$ is the survivor function at x for a non-central chi-squared random variable with ν degrees of freedom and non-centrality parameter λ, $y = kK^{2-\beta}$, $x = kS_t^{2-\beta} e^{r(2-\beta)\tau}$, and

$$k = \frac{2r}{\delta^2(2-\beta)(e^{r(2-\beta)\tau} - 1)}$$

(Randal 1998). The Black-Scholes formula is given as the special limiting case of the CEV solutions for $\beta < 2$ and $\beta > 2$ as we let $\beta \to 2$.

The compound option pricing formula (Geske 1979) introduces the debt payment M, with time to maturity $\tau_d > \tau$, i.e., greater than that of the option. In addition, stock price is a function of firm value V_t and the volatility of the firm's (homogeneous) assets σ. The option pricing formula is given by

$$C_t = V_t \Phi_2\left(h_1, h_2; \sqrt{\tfrac{\tau}{\tau_d}}\right) - Me^{-r\tau_d} \Phi_2\left(h_1 - \sigma\sqrt{\tau}, h_2 - \sigma\sqrt{\tau_d}; \sqrt{\tfrac{\tau}{\tau_d}}\right) - Ke^{-r\tau} \Phi(h_1 - \sigma\sqrt{\tau}) \tag{4.22}$$

where $\Phi_2(x, y; \rho)$ is the cumulative distribution function for the standard bivariate normal distribution at x and y with correlation ρ, $h_2 = g_t$, defined in (4.8) and

$$h_1 = \frac{\ln V_t - \ln \bar{V}_1 + (r + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}$$

where $\bar{V}_1$ is the value of V_t which solves $S(V_t, t) = K$ and where $S(V_t, t)$ is the function in (4.7).

The displaced diffusion option pricing formula (Rubinstein 1983) is relatively simple as a result of its assumptions. The call price is a function of firm value V_t, the value of the firm's non-risky assets $R_t = R_0 e^{rt}$, and the volatility of the risky assets σ, as well as the value at t of the risk-free debt B_t. The option price is

$$C_t = BS(V_t - R_t, K - \Delta_t e^{r\tau}, \tau, r, \sigma) \tag{4.23}$$

where $\Delta_t = R_t - B_t$, and $BS(\cdot)$ is the Black-Scholes equation defined in (4.20). In this case, the surplus risk-free assets are used to reduce the exercise payment of the call, and the GBM process is $A_t = V_t - R_t$ rather than the more familiar S_t or V_t. (Note also that $K > \Delta_t e^{r\tau}$ is assumed.)

The option pricing models (4.21), (4.22) and (4.23) can, given parameter inputs, be used to give call option prices in the same way as the Black-Scholes equation (4.20). They could equally be used to imply unknown parameters, such as β and δ in the CEV case, or R_0 and σ in the displaced diffusion model. If we artificially generate option prices using one of these alternative models, and then compute Black-Scholes implied volatilities, a volatility "smirk", monotonically increasing or decreasing in strike price, will result simply because the Black-Scholes model cannot adequately describe these prices. If we were to use the correct model to imply missing parameters, these will of course be constant.

The link between the volatility smirk (the monotonic relationship between implied volatility and strike price) and leverage, is rather tenuous. The volatility smirk is a function of the strike price of options held on the firm's stock; however the leverage effect concerns the behaviour of stock price volatility as a function of stock price. Nonetheless, it is common to think of a high strike price being consistent with a low stock price. This follows if we think about the payoff of a call option: when the strike price is high, the call option will be exercised if we see a large increase in the stock price, and we can consider the current stock price to be low.

The classical leverage effect is that, for high stock price, there is low stock price volatility, and for low stock price, high stock price volatility. If S_t is high, and volatility is indeed low, Black-Scholes prices the option using this low volatility. As a consequence, Black-Scholes prices are too low. Inferring σ using the Black-Scholes

formula (i.e. computing implied volatilities) based on the correct (higher) market call prices, we obtain a higher implied volatility than the correct value. High stock price is loosely equivalent to low strike price, and thus we see high implied volatilities for small K.

If S_t is low, and volatility is high, the opposite occurs. Black-Scholes prices according to this high volatility, and so Black-Scholes prices are too high. Implied volatility based on the lower market price is thus too low, and we see low implied volatilities for large K.

The underlying link between the volatility smirk and leverage effects is a complicated one, arising from the shape of the probability distribution for the stock price at exercise, which is of course related to current stock price volatility. More careful analysis of the cause of the volatility smirk is given in Appendix E for the displaced diffusion model.

As an illustration, the volatility smirk given by Hull (1997) for options on the S&P 500 Index on May 5, 1993 has a large negative gradient, with a range of approximately 85% to 160% of the at-the-money implied volatility (with $K = S_t$). These implied volatilities have been approximately transcribed from Hull's Figure 19.4 and reproduced in Figure 4.7. Using extreme parameter choices in the CEV model (4.21): $\beta = -20$ and $\delta = 0.11 S_t^{(2-\beta)/2}$, with $S_t = 100$, $\tau = 0.25$ and $r = 0.06$, we obtain call prices using (4.21), and the implied volatilities from (4.20) are then plotted in Figure 4.7 (this particular choice of δ implies index price volatility at t is 0.11). Although the parameters for the synthetic option prices have been chosen to obtain a range of values similar to those reported by Hull, rather than being motivated by time series properties of the index price, the similarities between the real implied volatilities (taken approximately from Hull's plot) and the synthetic smile in Figure 4.7 are striking. In particular, both smiles show an approximately linearly decreasing relationship between strike price and volatility, however the synthetic data does not lead to the same curvature for high K. Put another way, had Hull used market call prices on the index and the CEV model to imply parameters for the index price process, he would have obtained β and δ estimates very close to $\beta = -20$ and $\delta = 0.11 S_t^{(2-\beta)/2}$, and these would have been approximately constant across all options.

It does not appear to be possible to reproduce the smile shown in Figure 4.7 using the compound option model. Although compound option prices will imply volatilities

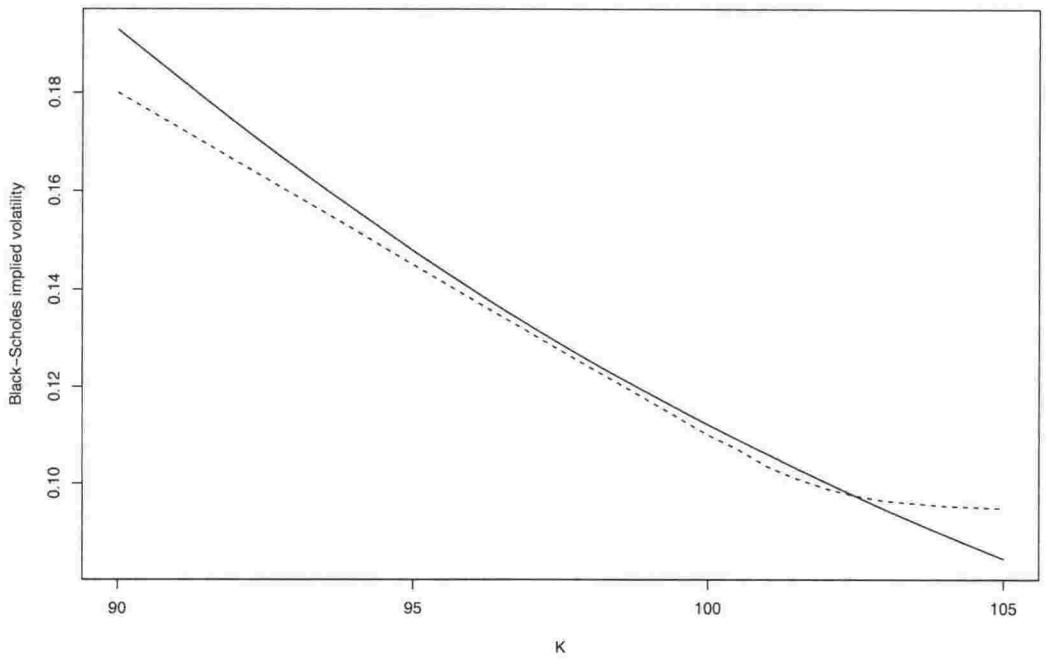

Figure 4.7. Black-Scholes implied volatilities for CEV option prices, given using the solid line, and an approximation to the smile given in Hull (1997) for the S&P 500 Index on May 5, 1993, given by the dotted line. The CEV option prices were computed using (4.21) with $S_t = 100$, $\tau = 0.25$, $r = 0.06$, $\beta = -20$ and $\delta = 0.11 S_t^{(2-\beta)/2}$ for the range of K shown, and the Black-Scholes implied volatilities numerically obtained via (4.20). The implied volatilities for the S&P 500 were read to a close approximation from Hull's Figure 19.4.

which are decreasing in K, the leverage effect is nowhere near strong enough to give such a steep smile. Since the CEV model is well defined mathematically for $\beta < 0$, we were able to choose $\beta = -20$ and obtain a very close match, however restricting $\beta \geq 0$ would have had similar implications to the use of the compound option model. The displaced diffusion model will produce a positive relationship between implied volatility and K, as would CEV with $\beta > 2$ and so neither of these models are consistent with the smile in Figure 4.7.

In conclusion, implied volatility is a poor forecast of future "realised" volatility because the incorrect option pricing formula is used to back out the volatility estimate. Acknowledgment that share prices do not typically follow GBM, and that as a consequence of this the *Black-Scholes* implied volatility is inappropriate, is important. If we can find an appropriate model for option prices, consistent with time-series properties of stock price for example, then volatility estimates implied from observed prices using this particular model will be highly desirable estimates for reasons discussed by Mayhew (1995).

4.2 The extended compound option pricing model

At this point we depart slightly from our focus on the underlying stock price processes assumed by existing option pricing models, to derive a new option pricing model. As we will see, this new model will afford us flexibility similar to that of the CEV model, in that we can model both an increasing or decreasing leverage effect within the one model.

In a recent textbook on option pricing, Hull (2000) considers analytical extensions to the Black-Scholes model, and makes mention of Geske's compound option model, and Rubinstein's displaced diffusion model, as well as the CEV model. Extensions to Geske (1979) and Rubinstein (1983) exist, but none address the union of the two approaches as we do here. Frey & Sommer (1998) discuss the extension of Geske (1979) to allow for both deterministic and stochastic interest rates. Chen & Ryan (1996) relax Rubinstein's (1983) assumptions to allow two classes of risky assets, rather than one risky and one non-risky; however like Rubinstein, they treat the fixed debt as risk-free. Toft & Prucyk (1997) obtain a call pricing formula in the presence of debt, but assume a different structure to either Rubinstein or Geske. In particular, using the model of Leland (1994), equity value is a deterministic function of debt characteristics, and only homogeneous assets are considered.

Here, using the results of Geske (1977), we combine the models of Geske (1979) and Rubinstein (1983), both of which recognise corporate debt, to produce an analytic formula for a call option over stock in a firm that has both debt and heterogeneous assets, without imposing any restriction to rule out bankruptcy. Unlike Geske, we allow the firm's assets to be heterogeneous. In particular, as in Rubinstein, some assets evolve according to GBM and the remaining assets are risk-free. Unlike Rubinstein, we allow for bankruptcy by interpreting the stock as a (compound) option over the value of the firm, as in Geske (1977, 1979). As a further extension to Geske's (1979) result, we allow the firm to have multiple debt repayments.

4.2.1 Call value with coupon bonds when V_t follows GBM

In order to generalise Geske's (1979) option pricing formula, we first obtain the formula for a compound option when the firm has an outstanding coupon bond, and firm value follows GBM.

Define

$$V_t = \text{the value of the firm at time } t;$$
$$S_t = \text{the value of the firm's stock at time } t; \text{ and}$$
$$B_t = \text{the value of the firm's coupon bond at time } t$$

and note that the coupon bond pays coupons $X_1, \ldots, X_{n-1}$ at times $t_1 < \cdots < t_{n-1}$, where $t_1 > t$, and a redemption payment M at time $t_n > t_{n-1}$. For ease of exposition, let $X_n = M$, and thus we can consider a general stream of coupons. The problem of finding B_t is addressed by Geske (1977), who gives the price of the coupon bond as a function of firm value. He makes the unnecessary assumption that the debt is repaid at constant intervals, with $t_i - t = i$ for each repayment. We provide an alternate proof of the result for arbitrary repayment dates in Appendix D.2. Using the relationship

$$S_t = V_t - B_t$$

and the formula for B_t given by Geske adapted to allow general payment dates, the value of the stock at time t is given by

$$G_n(V_t, \mathbf{X}, \boldsymbol{\tau}, r, \sigma) \equiv V_t \Phi_n(h_i; \{\rho_{ij}\}) - \sum_{m=1}^{n} X_m e^{-r\tau_m} \Phi_m(h_i - \sigma\sqrt{\tau_i}; \{\rho_{ij}\}) \qquad (4.24)$$

where $\mathbf{X} = (X_1, \ldots, X_n)$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$, r is the continuously compounding risk-free rate, σ is the (constant) volatility of firm value, $\Phi_n(h_i; \{\rho_{ij}\})$ is the cumulative distribution function of a standard n-variate normal random variable evaluated at upper limits $h_1, \ldots, h_n$ and with correlation matrix given by $\{\rho_{ij}\}$ for $1 \leq i, j \leq n$. Also, $\tau_i = t_i - t$ is the length of time until payment i, and

$$h_i = \frac{\ln V_t - \ln \bar{V}_i + (r + \frac{1}{2}\sigma^2)\tau_i}{\sigma\sqrt{\tau_i}}$$
$$\bar{V}_i = \begin{cases} \text{the value of } V \text{ which solves } S_{t_i}(V) = X_i & 1 \leq i \leq n-1 \\ X_n & i = n \end{cases}$$
$$\rho_{ij} = \sqrt{\frac{\tau_i}{\tau_j}} \qquad i < j$$

and $\rho_{ji} = \rho_{ij}$.

Derivation of this formula follows the suggestion of Black & Scholes (1973), who describe the common stock of a firm with a coupon bond as "an option on an option on ... an option on the firm". To see this, note that at any coupon time

	Stock		Call	
Time	Exercise price	Exercise choice	Exercise price	Exercise choice
t_1	X_1	$\max(S_{t_1} - X_1, 0)$	K	$\max(S_{t_1} - K, 0)$
t_2	X_2	$\max(S_{t_2} - X_2, 0)$	X_2	$\max(S_{t_2} - X_2, 0)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{n-1}	X_{n-1}	$\max(S_{t_{n-1}} - X_{n-1}, 0)$	X_{n-1}	$\max(S_{t_{n-1}} - X_{n-1}, 0)$
t_n	X_n	$\max(V_{t_n} - X_n, 0)$	X_n	$\max(V_{t_n} - X_n, 0)$

Table 4.1. Comparison of exercise prices and decisions faced by stockholders in Geske (1977), and a call option holder over a stock for a firm which has an outstanding coupon bond paying coupons $X_2, \ldots, X_n$.

t_i the stockholder can purchase a further option by paying the coupon X_i. Each option is ultimately an option over the assets of the firm. The formula (4.24) is derived by noting that, at the final debt payment, the value of the stock is $S_{t_n} = \max(V_{t_n} - X_n, 0)$, while at each earlier coupon payment, the stockholder chooses between default, and payment of the coupon, i.e. $S_{t_i^+} = \max(S_{t_i} - X_i, 0)$, where t_i^+ denotes the instant after time t_i. Using Itô's Lemma and a continuously rebalanced hedge portfolio, the partial differential equation for stock value does not feature investor preferences. Consequently risk-neutral pricing techniques are warranted, and a series of nested integrals may be evaluated using the multivariate normal probability functions featuring in (4.24).

In order to price a European call option on the stock, we redefine the coupon bond to pay coupons $X_2, \ldots, X_{n-1}$ at times $t_2 < \cdots < t_{n-1}$ and maturity payment $X_n = M$ at $t_n > t_{n-1}$. The call option over the stock has exercise price K, and if rational, this is made at time $t_1 < t_2$. Comparison of the exercise payments and decisions faced by the call-holder is made to those of the stockholder of Geske (1977), and this appears in Table 4.1. If we set $X_1 = K$, since the underlying stochastic process V_t is identical in each case, we must have $C_t = S_t$, where the latter is given by (4.24).

Thus, in order to value the call, we note that the call in this case matches the stock whose price is given in (4.24), with $X_1 = K$. Defining $X_1 = K$, the value of the call at time t is given by

$$C_t = G_n(V_t, \mathbf{X}, \boldsymbol{\tau}, r, \sigma). \tag{4.25}$$

This is equivalent to redefining the coupon dates of the firm, and pricing the call option with the first "coupon" equal to the exercise price of the call. This formula nests the Black-Scholes model, where no debt exists and hence $V_t = S_t$, with

$$C_t = G_1(S_t, K, \tau, r, \sigma) = S_t\Phi(h_t) - Ke^{-r\tau}\Phi(h_t - \sigma\sqrt{\tau})$$

where

$$h_t = \frac{\ln S_t - \ln K + (r + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}},$$

$\tau = \tau_1$, and univariate normal probabilities are used. The formula (4.25) also includes the compound option pricing model of Geske (1979), where $n = 2$, as a special case. For this single debt payment, call price is

$$C_t = G_2(V_t, (K, M), (\tau_1, \tau_2), r, \sigma)$$

where the debt M is repaid at time t_2, with time-to-maturity τ_2.

4.2.2 Call value with coupon bonds when V_t follows the displaced diffusion model

The displaced diffusion model was described in Section 4.1.3. In particular firm value at time t is defined via the equations

$$V_t = A_t + R_t \tag{4.26}$$

$$A_t = A_0 \exp\{(r - \tfrac{1}{2}\sigma^2)t + \sigma W_t\} \tag{4.27}$$

$$R_t = R_0 e^{r\tau}$$

for all $t > 0$, where $A_0 = \alpha_0 V_0$ is the initial investment in the risky assets, $R_0 = (1 - \alpha_0)V_0$ is the initial investment in risk-free assets, r is the continuously compounding risk-free rate, σ is the volatility of the risky assets, and W_t is a Brownian motion process under the risk-neutral measure $\mathbb{Q}$. Recall that $\alpha_t \equiv \frac{A_t}{V_t}$ is the proportion of firm value in the risky assets at time t, and by the properties of R_t, this satisfies

$$(1 - \alpha_t)V_t = (1 - \alpha_0)V_0 e^{rt} \tag{4.28}$$

for all t.

On the other side of the balance sheet, we have (as in the previous section) the firm's stock with value S_t, and debt, consisting of promised payments $X_2, \ldots, X_n$ at times $t_2 < \ldots < t_n$. We aim to value a call option on the stock, with exercise price K, and maturity at t_1.

Before presenting and proving the general result, it is useful to first work through the special case when $n = 2$, and the firm's debt is in the form of a discount bond maturing at time $t_2 > t_1$, where t_1 is the exercise date of the option.

Theorem 4.6 *The price at time t of a European call option over a stock with exercise price K payable at t_1, for a firm whose value evolves according to (4.26) and (4.27) and which has outstanding a single discount bond with redemption payment M due at time $t_2 > t_1$ is given by*

$$C_t = \begin{cases} G_2(A_t, (K, M - R_t e^{r\tau_2}), (\tau_1, \tau_2), r, \sigma) & R_t < Me^{-r\tau_2} \\ G_1(A_t, K + (Me^{-r\tau_2} - R_t)e^{r\tau}, \tau_1, r, \sigma) & Me^{-r\tau_2} \leq R_t < Ke^{-r\tau_1} + Me^{-r\tau_2} \\ V_t - Ke^{-r\tau_1} - Me^{-r\tau_2} & R_t \geq Ke^{-r\tau_1} + Me^{-r\tau_2} \end{cases} \tag{4.29}$$

where $\tau_i = t_i - t$ for $i = 1, 2$, $A_t = V_t - R_t$ is the value of the firm's risky assets at time t, and R_t is the value of the firm's risk-free assets at time t.

Proof Following discussion in both Geske (1979) and Rubinstein (1983), risk-neutral pricing is a valid way to price contingent claims for this firm. In particular, the stock price at time t_1 is given by

$$S_{t_1} = e^{-r(t_2 - t_1)} \mathrm{E}_{t_1}^{\mathbb{Q}} \{\max(V_{t_2} - M, 0)\} \tag{4.30}$$

where $\mathrm{E}_{t_1}^{\mathbb{Q}}$ is the expectation under the risk-neutral measure taken conditional on information available at time t_1. In calculating S_{t_1}, there are two cases to consider.

Case 1: If $R_t \geq Me^{-r\tau_2}$ the debt is risk-free. So,

$$S_{t_1} = V_{t_1} - Me^{-r(t_2 - t_1)} = A_{t_1} - (Me^{-r\tau_2} - R_t)e^{r\tau_1}$$

since $V_{t_1} = A_{t_1} + R_t e^{r\tau_1}$.

Case 2: If $R_t < Me^{-r\tau_2}$ there is a positive probability that the total value of the firm will not exceed M at t_2, and so we must evaluate the expectation (4.30). Following Rubinstein, we write

$$S_{t_1} = e^{-r(t_2 - t_1)} \mathrm{E}_{t_1}^{\mathbb{Q}} \{\max(A_{t_2} - (M - R_t e^{r\tau_2}), 0)\}$$

where we have written $V_{t_2} = A_{t_2} + R_t e^{r\tau_2}$. Rather than evaluating this integral directly, we note that A_{t_2} is a GBM process, and hence the solution is given by the Black-Scholes equation with exercise price $M - R_t e^{r\tau_2}$, i.e.

$$S_{t_1} = G_1(A_{t_1}, M - R_t e^{r\tau_2}, t_2 - t_1, r, \sigma).$$

Thus, at time t_1, the stock is worth

$$S_{t_1} = \begin{cases} G_1(A_{t_1}, M - R_t e^{r\tau_2}, t_2 - t_1, r, \sigma) & R_t < Me^{-r\tau_2} \\ A_{t_1} - (Me^{-r\tau_2} - R_t)e^{r\tau_1} & R_t \geq Me^{-r\tau_2} \end{cases} \tag{4.31}$$

where $A_{t_1} = V_{t_1} - R_{t_1}$ is the value of the risky assets at t_1.

We now consider pricing an option over the stock, maturing at time t_1, with exercise price K. Risk-neutral pricing applies and we must evaluate

$$C_t = e^{-r\tau_1} \mathrm{E}_t^{\mathbb{Q}}\{\max(S_{t_1} - K, 0)\} \tag{4.32}$$

where S_{t_1} is given by (4.31). In the case of non-risky debt, there are two possibilities.

Case A(i): The first of these is where the size of the non-risky assets ensures that $S_{t_1} > K$. If $R_t \geq Ke^{-r\tau_1} + Me^{-r\tau_2}$ both "options" will be exercised, and

$$C_t = V_t - Ke^{-r\tau_1} - Me^{-r\tau_2}.$$

Case A(ii): If the non-risky assets meet the debt payment (and ensure $V_{t_2} > M$), but are not large enough to also guarantee $S_{t_1} > K$, we have

$$C_t = e^{-r\tau_1} \mathrm{E}_t^{\mathbb{Q}}\{\max(A_{t_1} - [K + (Me^{-r\tau_2} - R_t)e^{r\tau_1}], 0)\}$$

and as before, we recognise that since A_{t_1} is a GBM process, the solution is given by the Black-Scholes equation with exercise price $K + (Me^{-r\tau_2} - R_t)e^{r\tau_1}$, i.e.

$$C_t = G_1(A_t, K + (Me^{-r\tau_2} - R_t)e^{r\tau_1}, \tau_1, r, \sigma).$$

Case B: When the debt is risky, i.e., $R_t < Me^{-r\tau_2}$, we must evaluate the integral

$$C_t = e^{-r\tau_1} \mathrm{E}_t^{\mathbb{Q}}\{\max[G_1(A_{t_1}, M - R_t e^{r\tau_2}, t_2 - t_1, r, \sigma) - K, 0]\}.$$

Again, rather than using brute-force, we note that Geske (1979) was faced with a similar integral

$$e^{-r\tau_1} \mathrm{E}_t^{\mathbb{Q}}\{\max(G_1(V_{t_1}, M, t_2 - t_1, r, \sigma) - K, 0)\}$$

where V_{t_1} was a GBM process. His solution was given by

$$G_2(V_t, (K, M), (\tau_1, \tau_2), r, \sigma)$$

in our notation, and hence the solution to the integral under the alternative conditions examined here is

$$C_t = G_2(A_t, (K, M - R_t e^{r\tau_2}), (\tau_1, \tau_2), r, \sigma).$$

Thus, combining the results for the risky and non-risky debt, we obtain the call pricing formula

$$C_t = \begin{cases} G_2(A_t, (K, M - R_t e^{r\tau_2}), (\tau_1, \tau_2), r, \sigma) & R_t < Me^{-r\tau_2} \\ G_1(A_t, K + (Me^{-r\tau_2} - R_t)e^{r\tau_1}, \tau_1, r, \sigma) & Me^{-r\tau_2} \leq R_t < Ke^{-r\tau_1} + Me^{-r\tau_2} \\ V_t - Ke^{-r\tau_1} - Me^{-r\tau_2} & R_t \geq Ke^{-r\tau_1} + Me^{-r\tau_2} \end{cases}$$

where $A_t = V_t - R_t$ as required. □

Note that the middle case of non-risky debt, but risky call exercise, was exactly that examined by Rubinstein (1983). Noting that $Me^{-r\tau_2}$ in the second case is the present value of the outstanding debt, we see that the two formulae are certainly consistent. Rubinstein does not consider the unlikely third case where exercise of the call option is guaranteed, nor does he provide a formula when the debt is risky.

We now present the general result for a firm with an outstanding stream of $n-1$ debt payments.

Theorem 4.7 (The extended compound option pricing model) *The price at t of a call option over a stock with exercise price X_1 payable at $t_1 < t_2$, for a firm whose value evolves according to (4.26) and (4.27) and which has outstanding debt with a stream of promised payments $X_2, \ldots, X_n$ due at times $t_2 < \cdots < t_n$ is given by*

$$C_t = \begin{cases} G_k(A_t, \mathbf{\Pi}_k, \boldsymbol{\tau}_k, r, \sigma) & k > 0 \\ V_t - \sum_{i=1}^{n} X_i e^{-r\tau_i} & k = 0 \end{cases} \tag{4.33}$$

where $A_t = V_t - R_t$ is the value of the risky assets of the firm at time t, $\tau_i = t_i - t$, $\boldsymbol{\tau}_k = (\tau_1, \ldots, \tau_k)$, G_k is given by (4.25) for $k > 0$,

$$\mathbf{\Pi}_k = \begin{pmatrix} X_1 \\ \vdots \\ X_{k-1} \\ \left(\sum_{i=k}^{n} X_i e^{-r\tau_i} - R_t\right) e^{r\tau_k} \end{pmatrix}$$

where R_t is the value of the non-risky assets of the firm at time t, and where k is chosen so that t_k is the earliest time at which the non-risky assets of the firm meet all subsequent debt payments if such a time exists, and n otherwise, i.e., k is the smallest non-negative integer that satisfies

$$R_t \geq \sum_{i=k+1}^{n} X_i e^{-r\tau_i} \tag{4.34}$$

if such a number exists and n otherwise.

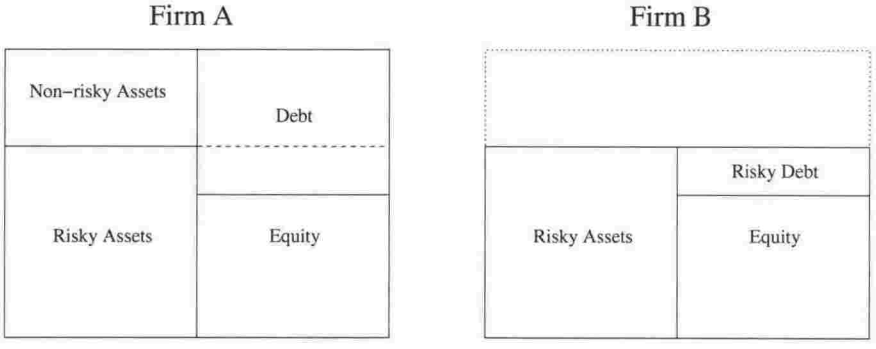

Figure 4.8. The balance sheets of two firms. Firm A has risk-free assets which offset some of the debt on the right hand side of the balance sheet. Firm B has only risky assets and a reduced amount of debt. For the purposes of valuing the firms' stock and European call options on the stock, the two firms are identical.

Before embarking on the proof to this theorem, it is useful to consider the intuition behind the result. In order to value the call, we compare the value of the non-risky assets to the present value of all outstanding exercise payments. We choose k so that all exercise payments (coupons) after X_k are met by the non-risky assets. The value of the surplus non-risky assets at time t_k is

$$0 < R_t e^{r\tau_k} - \sum_{i=k+1}^{n} X_i e^{-r(t_i - t_k)} \leq X_k$$

and thus the exercise payment at X_k is reduced by this amount. This remainder, and all earlier exercise payments, are made subject to available resources and limited liability. Having eliminated the non-risky assets, the call is then priced using the formula of Geske (1977), with first argument A_t (the process that follows GBM, and the residual assets of the firm) with exercise payments $X_1, \ldots, X_{k-1}$ and the reduced payment at t_k. At t_k all the subsequent payments are met by a fund consisting of only the risk-free assets, and any leftover risk-free assets are used to reduce X_k. Thus some of the debt on one side of the balance sheet is offset by the non-risky assets on the other side of the balance sheet. In this way, the non-risky assets are eliminated from the call pricing problem and we consider a matching firm, with a reduced quantity of debt, and risky assets following GBM. This situation is shown in Figure 4.8.

Having seen the intuition behind the theorem, we now consider its proof.

Proof First, we confirm the case where $n = 2$. This is proved in Theorem 4.6, so we need only confirm that (4.33) is identical to (4.29) when $n = 2$. We note that

$X_1 = K$ and $X_2 = M$ and that these payments are made at t_1 and t_2. If the value of the risk-free assets is sufficiently large that

$$R_t \geq \sum_{i=1}^{n} X_i e^{-r\tau_i}$$

then $k = 0$, ensuring that both the call will be exercised and the debt will be paid off, and hence the call value is given in (4.33) by the net present value

$$C_t = V_t - Ke^{-r\tau_1} - Me^{-r\tau_2}$$

which corresponds to the third case of (4.29) as required. If $R_t \geq X_2 e^{-r\tau_2}$ is true, but $R_t \geq X_1 e^{-r\tau_1} + X_2 e^{-r\tau_2}$ is not, then $k = 1$ and (4.33) specifies

$$\mathbf{\Pi}_1 = (Ke^{-r\tau_1} + Me^{-r\tau_2} - R_t)e^{r\tau_1}$$

and $\boldsymbol{\tau} = \tau_1$, corresponding to the second case of (4.29) as required. Finally, if $R_t < X_2 e^{-r\tau_2}$, no value of k satisfies (4.34) and so $k = 2$. In this case, (4.33) specifies

$$\mathbf{\Pi}_2 = \begin{pmatrix} K \\ (Me^{-r\tau_2} - R_t)e^{r\tau_2} \end{pmatrix} = (K, M - R_t e^{r\tau_2})$$

and $\boldsymbol{\tau}_2 = (\tau_1, \tau_2)$, and this corresponds to the first case of (4.29) as required. Thus Theorem 4.7 is confirmed for the case $n = 2$.

For general n, we note that if

$$R_t \geq \sum_{i=1}^{n} X_i e^{-r\tau_i}$$

then $k = 0$ and all "options" will be exercised, since stock price (or firm value for the last payment) will be greater than the exercise amount at every exercise date. Thus, when $k = 0$

$$C_t = V_t - \sum_{i=1}^{n} X_i e^{-r\tau_i} \tag{4.35}$$

as required.

At time t_1, either that exercise and all subsequent ones are risk-free, in which case the call value at t is given by (4.35), or the exercise is risky, and we value the call by

$$C_t = e^{-r\tau_1} \mathrm{E}_{t_1}^{\mathbb{Q}}\{\max(S_{t_1} - X_1, 0)\}.$$

If $k = 1$, then all subsequent payments are guaranteed by the non-risky assets of the firm, and

$$S_{t_1} = V_{t_1} - B_{t_1} = A_{t_1} - \left(\sum_{i=2}^{n} X_i e^{-r\tau_i} - R_t \right) e^{r\tau_1}$$

where B_{t_1} is the value of the outstanding (risk-free) debt, and as in the proof to Theorem 4.6, we see that

$$C_t = G_1(A_t, \mathbf{\Pi}_1, \tau_1, r, \sigma)$$

where $A_t = V_t - R_t$,

$$\mathbf{\Pi}_1 = K + \left(\sum_{i=2}^{n} X_i e^{-r\tau_i} - R_t \right) e^{r\tau_1} = \left(\sum_{i=k}^{n} X_i e^{-r\tau_i} - R_t \right) e^{r\tau_1}$$

since $k = 1$, and where $X_1 = K$ as required.

In the case where $k > 1$, we use the method of induction to complete the proof. Assuming Theorem 4.7 is true for $n - 1$ payments, at time $t' = t_1$, the stock of the firm is a compound option over the assets of a firm with risky assets worth $A_{t'}$ and $n - 1$ outstanding debt payments $X'_i = X_{i+1}$ to be made at times $t'_i = t_{i+1}$ $(i = 1, \ldots, n-1)$ as shown in Figure 4.9. Since $k > 1$ for the call option, the time at which the non-risky assets meet all remaining payments is $t'_{k-1} = t_k$ and so we define $q = k - 1 > 0$. Thus, the theorem states

$$S_{t'} = G_q(A_{t'}, \mathbf{\Pi}'_q, \boldsymbol{\tau}'_q, r, \sigma) \tag{4.36}$$

where $A_{t'}$ is a GBM process, $\tau'_i = t'_i - t'$, $\boldsymbol{\tau}'_q = (\tau'_1, \ldots, \tau'_q)$,

$$\mathbf{\Pi}'_q = \begin{pmatrix} X'_1 \\ \vdots \\ X'_{q-1} \\ (\sum_{i=q}^{n-1} X'_i e^{-r\tau'_i} - R_{t'}) e^{r\tau'_q} \end{pmatrix}$$

and where q is chosen so that $t'_q = t_k$ is the earliest time at which the non-risky assets of the firm meet all subsequent debt payments if such a time exists, and $n-1$ otherwise.

Guided by the two alternative definitions of the coupon schedule shown in Figure 4.9, we make appropriate replacements in (4.36) and find the value of the stock at time t_1 is

$$S_{t_1} = G_{k-1}(A_{t_1}, \mathbf{\Pi}^{(1)}_{k-1}, \boldsymbol{\tau}^{(1)}_{k-1}, r, \sigma) \tag{4.37}$$

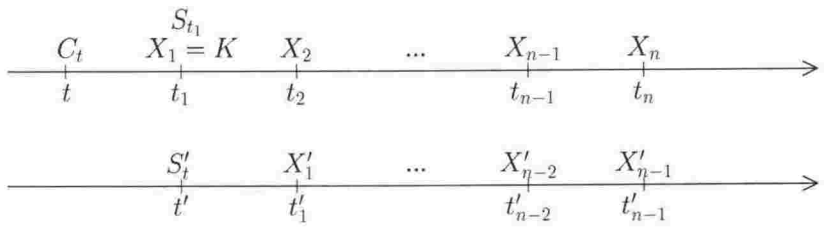

Figure 4.9. Two alternative representations of the coupon payments and payment dates relevant to the valuation of $S_{t_1} = S_{t'}$.

where A_{t_1} is a GBM process, the superscript (1) denotes pricing from time t_1,

$$\mathbf{\Pi}_{k-1}^{(1)} = \begin{pmatrix} X_2 \\ \vdots \\ X_{k-1} \\ (\sum_{i=k}^{n} X_i e^{-r\tau_i} - R_t)e^{r\tau_k} \end{pmatrix}$$

since from (4.28), $R_t = R_{t'}e^{-r(t'-t)}$, and where $\boldsymbol{\tau}_{k-1}^{(1)} = (t_2 - t_1, \ldots, t_k - t_1)$.

In order to value a compound option with m outstanding exercise payments when V_t follows GBM, following (4.25), an application of the result of Geske (1977) shows that its value at t is

$$C_t = G_m(V_t, \mathbf{X}_m, \boldsymbol{\tau}_m, r, \sigma) = e^{-r\tau_1}\mathrm{E}_{t_1}^{\mathbb{Q}}\{\max(S_{t_1} - X_1, 0)\} \tag{4.38}$$

where the stock price at t_1 is given by

$$S_{t_1} = G_{m-1}(V_{t_1}, \mathbf{X}_{m-1}^{(1)}, \boldsymbol{\tau}_{m-1}^{(1)}, r, \sigma), \tag{4.39}$$

and where V_{t_1} is the value of the GBM price process at t_1, $\mathbf{X}_{m-1}^{(1)} = (X_2, \ldots, X_m)$ and $\boldsymbol{\tau}_{m-1}^{(1)} = (t_2 - t_1, \ldots, t_m - t_1)$.

We see (4.39) and (4.37) are of identical form, but m, V_{t_1}, $\mathbf{X}_{m-1}^{(1)}$ and $\boldsymbol{\tau}_{m-1}^{(1)}$ in (4.39) are replaced by k, A_{t_1}, $\mathbf{\Pi}_{k-1}^{(1)}$ and $\boldsymbol{\tau}_{k-1}^{(1)}$ respectively in (4.37). Thus, in order to find the price of the call over stock whose value at exercise is given by (4.37) we replace m, V_t, $\mathbf{X}_m$ and $\boldsymbol{\tau}_m$ in (4.38) by k, $A_t = V_t - R_t$, $\mathbf{\Pi}_k$ and $\boldsymbol{\tau}_k$ respectively, where $\mathbf{\Pi}_k = (X_1, \mathbf{\Pi}_{k-1}^{(1)})$. Applying these changes to (4.38), in the presence of non-risky assets, the value of the call option is

$$C_t = G_k(A_t, \mathbf{\Pi}_k, \boldsymbol{\tau}_k, r, \sigma)$$

and this corresponds to (4.33) for $k > 0$. Thus, if $k > 1$, the theorem is true for n payments if it is true for $n - 1$ payments, and since the theorem is true for $n = 2$, by the process of induction, we conclude the theorem is true for all n. The cases $k = 0$ and $k = 1$ were previously shown true for general n, so Theorem 4.7 is true for all n and all values of k. □

4.2.3 Using the extended compound model to value coupon bonds

There are various special cases of the call pricing formula (4.33): the Black-Scholes equation, and Rubinstein's (1983) displaced diffusion option pricing model have already been mentioned, and correspond to the cases $n = 1$ with $X_1 = K$ and $R_t = 0$ for all t, for Black-Scholes, and $X_1 = K$, $n = 2$ with $X_2 = Me^{r\tau_2}$ with an assumption that $k \neq 2$ so that the debt is risk-free for Rubinstein. In addition, Geske's (1979) compound option pricing model is a special case with $X_1 = K$, $n = 2$, $X_2 = M$ and $R_t = 0$ for all t. In addition to these option pricing models, we can address the aims of Geske (1977), and use the model to price the stock of a firm with an outstanding coupon bond, and whose underlying firm value follows the displaced diffusion characterised by (4.27) and (4.26). Assuming $k \neq 0$, the value at t of the stock of such a firm, with outstanding coupon payments $\mathbf{X}$ to be made at times $t + \boldsymbol{\tau}$, is given by

$$S_t = G_k(A_t, \boldsymbol{\Pi}_k, \boldsymbol{\tau}_k, r, \sigma)$$

where $A_t = V_t - R_t$, k, $\boldsymbol{\Pi}_k$ and $\boldsymbol{\tau}_k$ are as specified in Theorem 4.7. Using the relationship $B_t = V_t - S_t$, the value of the outstanding coupons is given by

$$B_t = V_t\left(1 - \alpha_t \Phi_k(h_i'; \{\rho_{ij}\})\right) + \sum_{m=1}^{k} \Pi_m e^{-r\tau_m} \Phi_m(h_i' - \sigma\sqrt{\tau_i}; \{\rho_{ij}\})$$

where $\alpha_t = \frac{A_t}{V_t}$, Π_j is the jth element of $\boldsymbol{\Pi}_k$, and the h_i' are defined by amending their earlier definition in the obvious way. Once again, we note that Geske's (1977) assumption of regular debt repayments was unnecessary, as proved in Appendix D.2.

4.2.4 Some numerical results

Here we analyse call prices computed using (4.33). The multivariate normal probabilities are evaluated using the algorithm of Genz (1992) implemented in the statistical software R (Ihaka & Gentleman 1996). Genz's algorithm was compared to other algorithms by Genz (1993) and found to be the most efficient method of those considered.

Rather than comparing the call prices themselves (as in Rubinstein 1983), we use Black-Scholes implied volatilities to standardise the calls for changing exercise price. The ratios of those implied volatilities to the at-the-money implied volatility (where

$K = S_t$) for that firm are given in Table 4.2 in a format similar to that of Toft & Prucyk (1997). Each "firm", consisting of a leverage ratio and $\alpha_t = \frac{A_t}{V_t}$ combination, has $S_t = 10$ and $\sigma(S_t, t) = 0.4$ as described below, and we price the options at time t. By specifying the firm characteristics via the stock price rather than firm value, we focus attention on readily available data, and ensure fairer comparison between

payment(s) and σ to satisfy the time t constraints we impose. Thus, at time t, σ is given by

$$\sigma = \tilde{\sigma}(S_t, t)\frac{S_t}{A_t \frac{\partial S_t}{\partial V_t}}. \tag{4.40}$$

This mirrors the approach used by MacBeth & Merville (1980) in their analysis of the constant elasticity of variance model. Although the stock price processes will have the same volatility at t, the volatility at maturity of the option $(t + \tau_1)$, or average volatility over the remaining life of the option will not necessarily be equal. Alternative approaches have been used to align processes: Rubinstein (1983) matches $\text{var}_t\{\ln(S_T/S_t)\}$ for all processes considered, and Beckers (1983) matches $\text{var}_t(S_T/S_t)$, where $T = t + \tau_1$ is the exercise date of the options used to find implied volatilities. In this situation, either of these alternative approaches would be difficult to implement, and so we settle for matching the instantaneous volatilities of the processes at the beginning of the period of interest.

In order to determine σ, we need to evaluate $\frac{\partial S_t}{\partial V_t}$ and substitute into (4.40). When the firm has no debt, $S_t = V_t$ and so

$$\sigma = \frac{\tilde{\sigma}(S_t, t)}{\alpha_t} \tag{4.41}$$

where $\alpha_t \equiv \frac{A_t}{V_t} = \frac{A_t}{S_t}$ in the case of no debt.

In the case where the firm has outstanding debt payments $\mathbf{X}$ due at times $\boldsymbol{\tau}$, S_t is the price of a compound option and is given by (4.33), where in particular, this and $\frac{\partial S_t}{\partial V_t}$ depend on the number of outstanding risky payments k. When $k = 0$, no debt payments are risky, and the stock price is given by

$$S_t = V_t - \sum_{i=1}^{n} X_i e^{-r\tau_i}$$

with $\frac{\partial S_t}{\partial V_t} = 1$. Substituting into (4.40), we find

$$\sigma = \tilde{\sigma}(S_t, t)\frac{V_t - \sum_{i=1}^{n} X_i e^{-r\tau_i}}{A_t} \tag{4.42}$$

where $A_t = V_t - R_t$.

When $k > 0$, stock price is given by

$$S_t = G_k(A_t, \boldsymbol{\Pi}_k, \boldsymbol{\tau}_k, r, \sigma)$$

in the notation of Theorem 4.7. Generally, we obtain $\frac{\partial S_t}{\partial V_t}$ from this and substitute into (4.40). For the case $k = 1$, the stock price is given by the Black-Scholes formula with first argument A_t. In this case, $\frac{\partial S_t}{\partial A_t}$ is the well known hedge ratio and is

$$\frac{\partial S_t}{\partial A_t} = \Phi\left(\frac{\ln A_t - \ln \Pi_1 + (r + \frac{1}{2}\sigma^2)\tau_1}{\sigma\sqrt{\tau_1}}\right) \equiv \tilde{\Phi}_1(A_t, \Pi_1, \tau_1, r, \sigma).$$

Substituting into (4.40), and noting from (4.26) that for fixed t, $\frac{\partial S_t}{\partial V_t} = \frac{\partial S_t}{\partial A_t}$, we find

$$\sigma = \tilde{\sigma}(S_t, t)\frac{G_1(A_t, \Pi_1, \tau_1, r, \sigma)}{A_t\tilde{\Phi}_1(A_t, \Pi_1, \tau_1, r, \sigma)} \tag{4.43}$$

where $A_t = V_t - R_t$. When $k = 2$, the stock price is given using the call pricing formula of Geske (1979). He gives the derivative $\frac{\partial S_t}{\partial A_t}$ for the compound option with two outstanding (risky) debt payments (Geske 1979, equation 10) as

$$\frac{\partial S_t}{\partial A_t} = \Phi_2\left(g_1, g_2; \sqrt{\frac{\tau_1}{\tau_2}}\right) \equiv \tilde{\Phi}_2(A_t, \mathbf{\Pi}_2, \boldsymbol{\tau}_2, r, \sigma)$$

where

$$g_1 = \frac{\ln A_t - \ln \bar{V} + (r + \frac{1}{2}\sigma^2)\tau_1}{\sigma\sqrt{\tau_1}}, \qquad g_2 = \frac{\ln A_t - \ln \Pi_2 + (r + \frac{1}{2}\sigma^2)\tau_2}{\sigma\sqrt{\tau_2}}$$

and $\bar{V}$ satisfies $S_{t_1}(\bar{V}) = G_1(\bar{V}, \Pi_2, \tau_2 - \tau_1, r, \sigma) = \Pi_1$. Substituting into (4.40), and again noting $\frac{\partial S_t}{\partial V_t} = \frac{\partial S_t}{\partial A_t}$, we find

$$\sigma = \tilde{\sigma}(S_t, t)\frac{G_2(A_t, \mathbf{\Pi}_2, \boldsymbol{\tau}_2, r, \sigma)}{A_t\tilde{\Phi}_2(A_t, \mathbf{\Pi}_2, \boldsymbol{\tau}_2, r, \sigma)} \tag{4.44}$$

where $A_t = V_t - R_t$.

Equations (4.41) to (4.44) allow us to determine σ for call option valuation for up to two risky debt payments, with the additional risky exercise of the option. When $k > 2$, equivalent equations follow in the same manner, with the general form of $\frac{\partial S_t}{\partial V_t}$ derived in Appendix D.2. For one or more risky debt payments, numerical solution of (4.40) will be necessary, since σ features on the right hand side through both S_t and the normal probabilities.

In order to make this alignment procedure clearer, we provide an example for the case where $\alpha_t = 0.75$ and $LR_t = 0.5$, and we have only a single debt payment. Since $LR_t = 0.5$ and $S_t = 10$, we find $V_t = 20$, and consequently $A_t = 0.75(20) = 15$. We now have two unknowns, the promised debt payment X_1 and the volatility of the

firm's risky assets σ. If the debt were risk-free, we would set $X_1 = 10e^{r\tau}$, however this is not the case, and we must solve

$$10 = BS(15, X_2 - 5e^{r\tau_2}, \tau_2, r, \sigma) \tag{4.45}$$

$$0.4 = \left(\tfrac{15}{10}\sigma\right)\tilde{\Phi}_1(15, X_2 - 5e^{r\tau_2}, \tau_2, r, \sigma) \tag{4.46}$$

where (4.45) is the stock price requirement, (4.46) is the volatility requirement, $r = 0.05$, $\tau_2 = 2$, and X_2 and σ are unknown. These non-linear simulataneous equations are solved numerically to yield

$$X_2 = 11.05358 \quad \text{and} \quad \sigma = 0.266926.$$

Substituting these values back into (4.45) and (4.46), we confirm the solution. Note that since $X_2 e^{-r\tau_2} > B_t = 10$, the debt is not risk-free.

Having provided a method for aligning "firms", we can now analyse call prices computed from (4.33). Three debt schedules are considered. The first of the three column blocks in Table 4.2 gives implied volatility ratios for a firm with no debt. The first sub-column of this block, with $\alpha_t = 1$, corresponds to the Black-Scholes situation, and hence has constant implied volatility and unit ratios. The second block is for firms with a single outstanding debt payment. The size of this payment is determined by the leverage ratio, and is found to satisfy $S_t = 10$. The time-to-maturity of this payment is $\tau_2 = 2$. The first column of this block, with $\alpha_t = 1$, corresponds to Geske's (1979) compound option prices. The final block in the table is for firms with two outstanding debt payments. These are of equal size, again determined to satisfy the leverage ratio and $S_t = 10$, and have times-to-maturity $\tau_2 = 1$ and $\tau_3 = 1.5$. In every case, the continuously compounding risk-free rate is $r = 0.05$.

Several interesting features are evident in Table 4.2. We note that in each block (corresponding to a leverage ratio, and a debt schedule) there is a column of ones. This indicates a situation where the Black-Scholes formula is the appropriate pricing model. To the left of the column of ones, there is evidence of the classical leverage effect, where the Black-Scholes formula overprices out-of-the-money calls (i.e. the true call price has a low implied volatility), and underprices in-the-money calls (the true call price has a high implied volatility); an effect discussed in Section 4.1.4.

At the column of ones, the debt is met exactly by the non-risky assets, and the calls are priced by the Black-Scholes formula giving constant implied volatilities.

	No debt payment				Single debt payment				Two debt payments			
α_t	1	0.75	0.5	0.25	1	0.75	0.5	0.25	1	0.75	0.5	0.25
Strike	Leverage = 25%											
8	1.000	0.958	0.868	0.489	1.030	1.000	0.936	0.709	1.030	1.000	0.936	0.709
9	1.000	0.981	0.942	0.807	1.014	1.000	0.971	0.878	1.014	1.000	0.971	0.878
10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
11	1.000	1.016	1.048	1.145	0.988	1.000	1.024	1.096	0.988	1.000	1.024	1.096
12	1.000	1.029	1.087	1.261	0.978	1.000	1.044	1.174	0.978	1.000	1.044	1.174
	Leverage = 50%											
8					1.056	1.039	1.000	0.868	1.060	1.040	1.000	0.868
9					1.026	1.018	1.000	0.942	1.027	1.018	1.000	0.942
10					1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
11					0.977	0.984	1.000	1.048	0.976	0.984	1.000	1.048
12					0.957	0.971	1.000	1.087	0.956	0.971	1.000	1.087
	Leverage = 75%											
8					1.077	1.071	1.056	1.000	1.087	1.078	1.060	1.000
9					1.036	1.033	1.026	1.000	1.040	1.036	1.027	1.000
10					1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
11					0.968	0.971	0.977	1.000	0.965	0.969	0.976	1.000
12					0.939	0.945	0.957	1.000	0.934	0.941	0.956	1.000

Table 4.2. The ratio of Black-Scholes implied volatilities to the at-the-money implied volatility, for call options with time to maturity $\tau = 0.5$. All firms have $S_t = 10$ and $\sigma(S_t, t) = 0.40$. The single debt payment is at $\tau_2 = 2$, and the two debt payments are of identical size and made at $\tau_2 = 1$ and $\tau_3 = 1.5$. The leverage figure determines V_t, and this and $S_t = 10$ are used to find the required debt payment(s). The risk-free rate is $r = 0.05$ throughout.

This is an artifact of the choices of leverage ratio and α_t, e.g., when the leverage ratio is 75%, $V_t = 40$ and $B_t = 30$, since $S_t = 10$. Thus, when $\alpha_t = 0.25$, the non-risky assets, with value $R_t = 30$, exactly offset the debt, and Black-Scholes is the appropriate call valuation formula.

To the right of the column of ones, there is evidence of an opposite effect resulting from the non-risky assets of the firm. In the first block, in the absence of debt, as the proportion of risky assets in the firm falls below one, we see empirically atypical behaviour, namely implied volatilities increasing with strike price. This phenomenon is consistent with the call prices given in Rubinstein (1983), and is a complicated combination of two effects: the presence of risk-free assets requiring an increase in the volatility of the risky assets (to maintain a fixed volatility for stock price), and the risk-free assets offsetting the strike price of the option at maturity. This effect is further investigated in Appendix E.

We note that, when the leverage ratio is low, increasing the number of debt payments from one to two has resulted in identical Black-Scholes implied volatility ratios to the accuracy provided. This reflects the way the stock price processes have been aligned in each case.

For both the single debt payment, and two debt payments, each leverage effect magnifies as leverage increases. Further, for a given leverage ratio, each effect magnifies

as we increase the number of debt payments. Extending this behaviour, we conclude that a steeply decreasing relationship between Black-Scholes implied volatility and call option strike price (as seen in Hull (1997), page 504, for the S&P 500 Index) is consistent with a firm that is highly levered and whose debt repayment schedule consists of many individual payments. Of course, this is a natural description of firms in practice.

4.2.5 Properties of volatility and elasticity

Having derived a new option pricing model, we now turn to the properties of the underlying stock price process; in particular, its volatility and elasticity of volatility with respect to stock price.

We investigate the behaviour of volatility and the elasticity of this volatility, when firm value follows the displaced diffusion model, and debt is not assumed risk-free. As a result, the value of the stock of this firm is a compound option, and in particular, S_t is given by Theorem 4.7. The case where $k = 0$ is a risk-free debt model, and hence the behaviour of volatility and its elasticity are consistent with the displaced diffusion model given described in Section 4.1.3. If $k = 1$ and $\alpha_t = 1$, the model for stock price is the same as for the compound option pricing model, and is described in Section 4.1.2. Thus, we focus on the simplest remaining cases: where $k = 1$ and $\alpha_t < 1$, consistent with a firm with heterogeneous assets, and a single risky debt payment; and where $k = 2$ and $\alpha_t = 1$, consistent with a firm with homogeneous (risky) assets, and two risky debt payments.

When $k = 1$, and a single debt payment M is made at t_2, the stock price is given by $S_t = G_1(A_t, M - (1 - \alpha_t)V_t e^{r\tau_2}, \tau_2, r, \sigma)$ where $G_1(\cdot)$ is the Black-Scholes equation, and $\tau_2 = t_2 - t$. As with the displaced diffusion model, we fix the value of the risk-free assets R_t and allow A_t only to vary with V_t. With reference to (4.43), the volatility of S_t is given by

$$\sigma_S(V_t, t) = \sigma \frac{(V_t - R_t)\Phi(\gamma_t)}{(V_t - R_t)\Phi(\gamma_t) - \Pi_1 e^{-r\tau_1}\Phi(\gamma_t - \sigma\sqrt{\tau_2})} \tag{4.47}$$

where A_t is replaced by $V_t - R_t$, $\Pi_1 = M - R_t e^{r\tau_2}$, and

$$\gamma_t = \frac{\ln(V_t - R_t) - \ln \Pi_1 + (r + \frac{1}{2}\sigma^2)\tau_2}{\sigma\sqrt{\tau_2}}.$$

We note that this volatility has an identical functional form to the volatility for the compound model given in (4.9), but with $A_t = V_t - R_t$ as the leading argument,

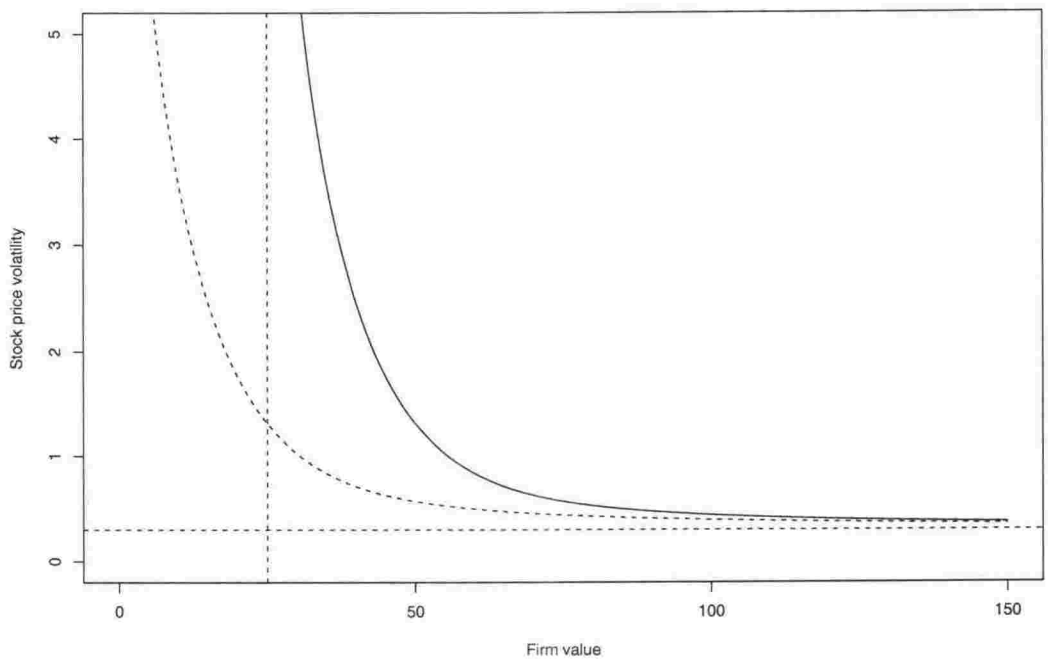

Figure 4.10. Stock price volatility for the extended compound model. The solid function is the volatility for a firm with $R_t = 25$, $\Pi_1 = 25$ due in $\tau_2 = 1$, $r = 0.05$ and $\sigma = 0.3$, and is given by (4.47). This function has the vertical asymptote at $V_t = R_t$ and horizontal asymptote at σ, and these are shown by the dashed line. The remaining function is for a firm with no risk-free assets with $X_1 = 25$ due at $\tau_1 = 1$, and is given by (4.9).

rather than V_t. Thus, the properties of the volatility for the extended compound model with $k = 1$ are identical to those of the compound model, as are the properties of the elasticity, except that the lower bound for V_t is now at R_t rather than zero.

An example of the volatility function (4.47) for a firm with $R_t = 25$, $\Pi_1 = 25$ due in $\tau_1 = 1$, $r = 0.05$ and $\sigma = 0.3$ is shown in Figure 4.10, along with the volatility for a comparable compound option process. It is clear from the plot that the extended compound volatility is just a translation of the compound function analysed in Section 4.1.2. In particular, the asymptote as firm value gets small is now at $V_t = R_t$, the value of the risk-free assets, rather than at $V_t = 0$. We also note that the compound option volatility is bounded between the extended compound volatility and the (constant) volatility of the risky assets for $V_t > R_t$.

The second case we consider is when a firm has two risky debt payments, and homogeneous assets (i.e. $\alpha_t = 1$). In this case, stock price volatility is given by

$$\sigma_S(V_t, t) = \sigma \frac{V_t}{S_t(V_t, t)} \tilde{\Phi}_2(V_t, \mathbf{X}, \boldsymbol{\tau}, r, \sigma) \tag{4.48}$$

where $S_t(V_t, t) = G_2(V_t, \mathbf{X}, \boldsymbol{\tau}, r, \sigma)$, $\mathbf{X} = (X_1, X_2)$ are the debt payments, and $\boldsymbol{\tau} = (\tau_1, \tau_2)$ the times-to-maturity of those debt payments.

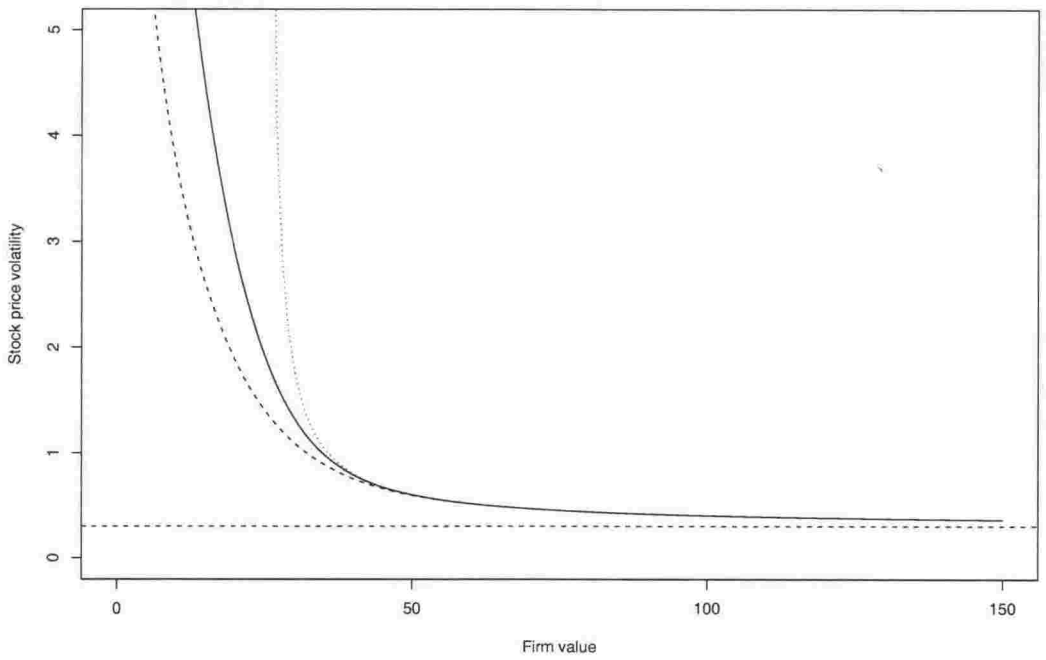

Figure 4.11. Stock price volatility for the extended compound model with no risk-free assets and two risky debt payments. The solid function is the volatility for a firm with $\mathbf{X} = (12.5, 12.5)$, $\boldsymbol{\tau} = (0.5, 1.5)$ and $\sigma = 0.3$, and is given by (4.48). This function has the vertical asymptote at $V_t = 0$ and horizontal asymptote at σ, and the latter is shown by the dashed line. The function shown by the dashed line is for a firm with no risk-free assets and a single debt payment $X_1 = 25$ due at $\tau_1 = 1$, and is given by (4.9). The third function, shown by the dotted line, is for a firm with risk-free debt, and $X_1 = 25$ due at $\tau_1 = 1$. The vertical asymptote for this function is at $V_t = 25$ and is not shown.

An example of the volatility function (4.48) for a firm with $X_1 = X_2 = 12.5$ with times-to-maturity $\tau_1 = 0.5$ and $\tau_2 = 1.5$ respectively, and $\sigma = 0.3$ is shown in Figure 4.11, plotted against firm value V_t. Also shown for comparison are the volatility for a firm with a single debt payment $X_1 = 25$ with time-to-maturity $\tau_1 = 1$, both under the risky compound model (4.9) and the risk-free model (4.3). In order to eliminate time-value issues, we set $r = 0$ so that the risk-free value of all debt schedules is $B_t = 25$. It is clear from the plot that the compound volatility with $k = 2$ (shown using the solid line) is very similar in behaviour to the volatility with $k = 1$ and the same amount of outstanding debt (shown using the dashed line). When firm value is high, and default unlikely, the volatility functions are indistinguishable; however, as firm value decreases, while both volatility functions become explosive, the volatility for $k = 2$ increases more rapidly than when $k = 1$.

From Figures 4.10 and 4.11, it appears that we can make the following generalisations. As firm value decreases, when we increase the number of debt payments, it appears that share value becomes increasingly volatile, reflecting transfer of risk

from debtholders to stockholders. The volatility under the assumption of risk-free debt is an upper bound for the volatility function as we increase k, with the limit representing the case where no risk is borne by the debtholders. It also appears that the elasticity of stock price volatility with respect to stock price when $k > 1$ is bounded by the elasticity functions for the Geske (1979) model (with $k = 1$) shown in Figure 4.3 and the risk-free debt model. Thus the relationship between the volatility functions observed in Figure 4.11 is consistent with the expected relationship. Introduction of risk-free assets to the firm serves only to provide a positive lower bound for firm value and shift the vertical asymptote of stock price volatility accordingly, as seen in Figure 4.10.

4.3 Data analysis

Using the iterated t-volatility estimator of Definition 3.3 to estimate volatility, we briefly analyse a single New Zealand stock, with a view to identifying plausible models from those outlined earlier in this Chapter. We focus on the CEV elasticity relationship, i.e., the relationship between log stock price ($\ln S_t$) and log volatility ($\ln \hat{\sigma}_t$). The gradient of the relationship between these variables is the elasticity, and for the CEV model, this gradient should be constant for all S_t (implying a linear relationship). Although we do not specifically match the non-linear compound and displaced diffusion relationships to what is observed, we do focus on the sign of the elasticity. This is facilitated by a non-parametric estimate of the relationship between the two variables obtained using `loess` (discussed in Appendix A). This non-parametric estimate is also a basis of comparison for the estimated linear relationship, and can be used to appraise whether the CEV model is appropriate for the data.

4.3.1 Analysis for Telecom NZ

We choose to analyse daily closing price data for Telecom Corporation of New Zealand Ltd. (Telecom) over the ten year period 20 March 1992 to 22 March 2002. Telecom is one of New Zealand's largest companies, and as a telecommunications company, is a prime candidate for the extended compound model. In particular, the assets of telecommunications companies: the relatively risk-free assets dedicated to fixed-line telephone services provision; and the high risk assets that characterise the

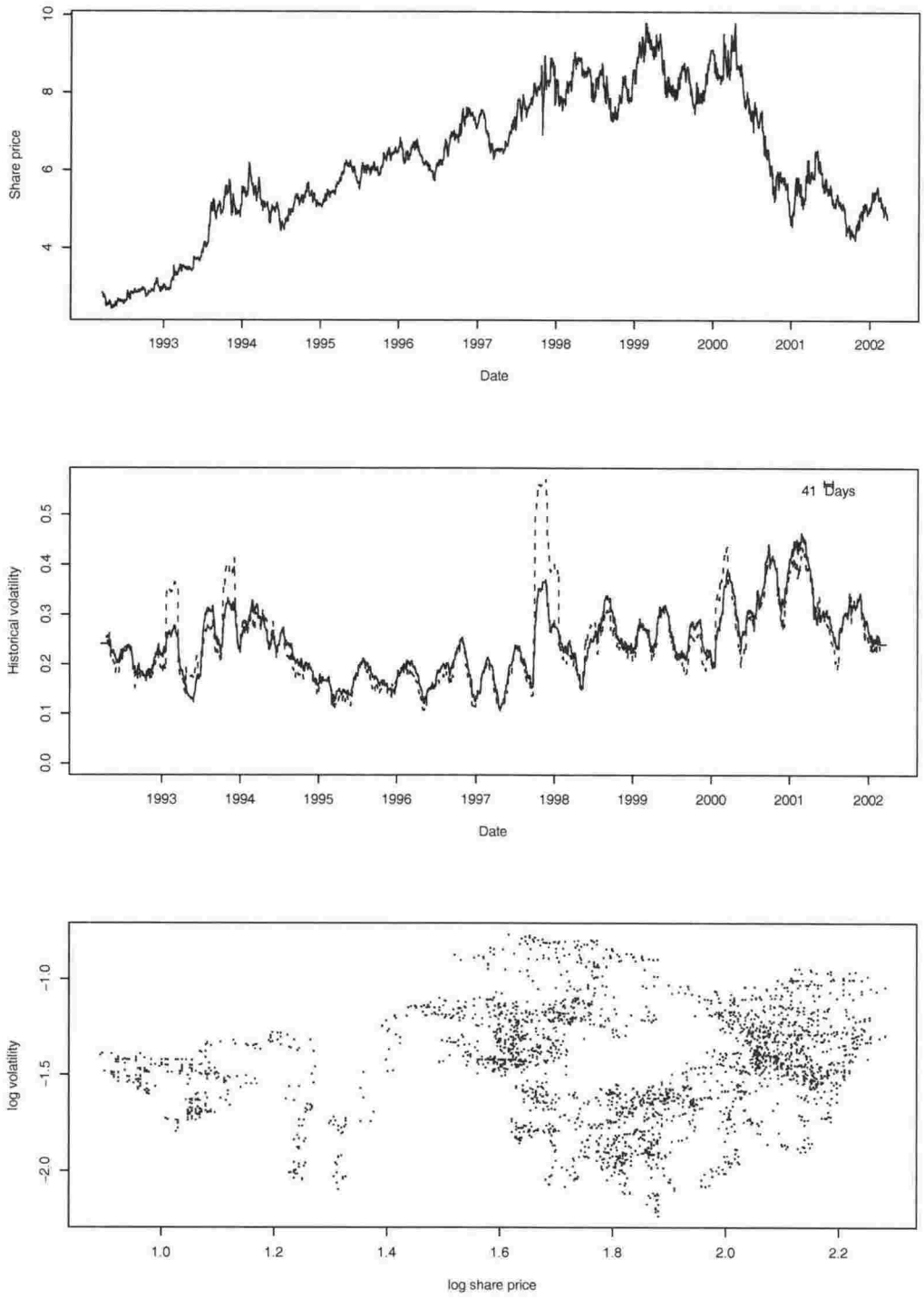

Figure 4.12. Telecom Corporation of New Zealand share price, volatility, and elasticity relationship. The top plot shows the share price series for the trading days in the period 20 March 1992 to 22 March 2002, where the labels mark the beginning of the year. The second plot gives the volatility estimates obtained using the iterated t-volatility estimator (using the solid line) and the moving standard deviation (using the dashed line) with a window width of 41 observations. The third plot shows the log of iterated t-volatility estimates plotted against log share price.

industry (including cellular phone networks, internet service provision, electronic-business applications and more), can be closely approximated by the heterogeneous asset decomposition of Rubinstein (1983).

The price series for Telecom for the entire period is shown in Figure 4.12. In addition, the volatility for this series is calculated using both the non-robust historical volatility estimator defined in (3.2) and the preferred estimator of Chapter 3 based on the t-distribution, with $\nu = 5$ degrees of freedom, as defined in Definition 3.3. As with the plots in Chapter 3, we see periods where the traditional, non-robust volatility estimate is unduly affected by long tails in the data.

Also shown in Figure 4.12, is a plot of log volatility (estimated using the iterated t-volatility estimator) against log stock price. If the CEV model prevails, we would expect the relationship between these two series to be linear. In order to satisfy the basic leverage arguments, the slope coefficient should be between -1 and 0 although this is not essential for use of the CEV model. Generally, a negative relationship wil be consistent with risky debt models, or CEV with $\beta < 2$, whereas a positive relationship will be consistent with the displaced diffusion model (with heterogeeous assets and non-risky debt), or CEV with $\beta > 2$. If GBM is appropriate, we would expect log volatility to be a constant linear function of log stock price, i.e. the relationship is a linear one, with zero slope. Acknowledging debt in the firm, we choose not to estimate the relationship between log volatility and log stock price in Figure 4.12, due to the high likelihood that the firm will have undergone capital structure changes over that ten year period, with the implication that no single relationship would apply for the whole period. In fact, the scatterplot in Figure 4.12 shows no discernible pattern, which is not surprising given ten years' data are shown.

The extended compound model presented in the previous section has several benefits over existing models for stock price: multiple debt repayments are allowed, and these are not assumed to be risk-free; and heterogeneous assets are acknowledged, and modelled in the form of risky assets whose value follows GBM, and risk-free assets whose value compounds at the risk-free rate. By modelling the firm in this way, we ultimately construct an option pricing formula which contains many arguments, but unlike the CEV model (for example), many of these arguments should be estimated from firm properties, rather than inferred, or estimated directly, using stock or option price data. In contrast, the CEV model features two unknown parameters which

must be estimated from time series stock price data, or implied using stock and option price data, but which do not have any direct meaning in terms of the firm's capital structure. It is difficult to motivate regular changes in these parameters; however, it is much less difficult to motivate changes in future debt schedules, and asset mix for the firm, since analysis of financial statements and general knowledge of capital structure reveals that these variables change as a matter of course.

We analyse the data for Telecom over periods of a single year for leverage effects, using the volatility estimate shown in Figure 4.12 estimated using the entire price series. Use of the estimate from Figure 4.12, rather than directly estimating volatility for each annual period, minimises both the problem of rescaling the volatility estimate using the sample variance, and end-effects. The first of these problems was discussed in Chapter 3, and we saw there that the sample variance for as many as 250 standardised returns can still be grossly inflated by a small number of extreme returns. Use of the volatility estimate based on all the data lessens this effect, since the series is ten times as long, and also means we lose observations only at the ends of the complete ten-year series, rather than at the ends of each year long sub-series.

We assume that over a single calendar year, the debt and asset mix parameters remain fixed, and that any plot of $\ln \hat{\sigma}_t$ against $\ln S_t$ is both resistant to actual changes in these values, and also to any dependence of σ_t on time (e.g. through a compound model). In particular, we hope that these effects are secondary to the dominant leverage effect.

Figure 4.13 features the plot of log volatility against log share price for Telecom for the nine calendar years 1993 to 2001. Added to each plot are the ordinary least squares regression line, and the robust non-paramtric relationship estimated using `loess` with a smoothing window of $\frac{1}{3}$. Except for those in years 1994, 1995, and 1999, all slope coefficients in Figure 4.13 are significantly different from zero at the 1% level, however the 1999 slope is significant at the 5% level. This provides evidence that the Telecom share price does not follow GBM (with a slope of zero), and also clear evidence that the slopes are not all equal, since significant negative, and positive slopes arise.

A linear relationship does seem plausible for many of the individual years: R^2 figures range from 0.2% in 1995 to 49.2% in 1997, corresponding to (absolute) linear correlation coefficients between 0.046 and 0.701. In many cases, the non-parametric relationship provided by `loess` does not depart greatly from the regression line for

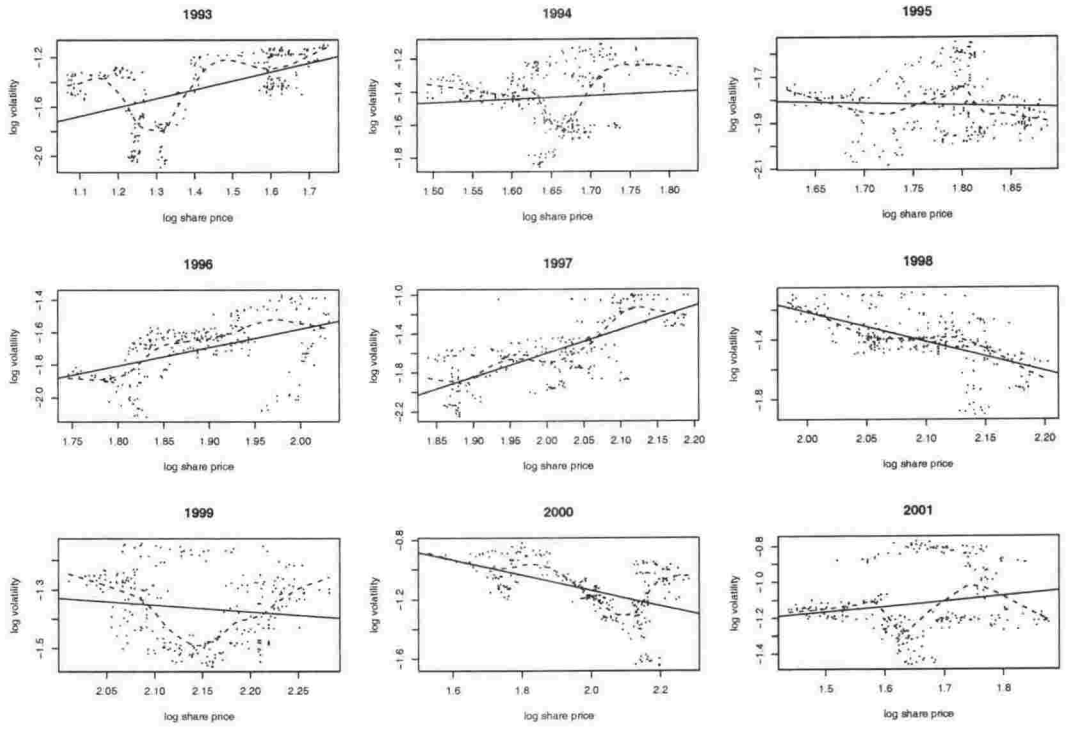

Figure 4.13. Leverage plots for Telecom New Zealand for 1993 to 2001. Volatility is estimated using Theorem 3.1 with $\nu = 5$ for the entire period, and log volatility is plotted against log price for each year-long period. Superimposed are the relationships estimated using linear regression, and the robust, non-parametric smoother `loess`.

that year. This supports the use of the CEV model for each of these periods; however the fact that the slope coefficients are not stable implies a time-varying elasticity parameter β. This is not ideal, and we would conclude that the CEV model is not suitable for long-term modelling of the stock price. Nonetheless, given the flexibility of the CEV to model both the classical leverage effect with $\beta < 2$ and an increasing relationship between log volatility and log stock price with $\beta > 2$, and the apparently linear relationship between log volatility and log price over the period of each year, pricing of short maturity options using the CEV model could be appropriate.

4.3.2 Reconciliation with stock price models

One implication of Figure 4.13 is that the compound or displaced diffusion models alone cannot be used to model Telecom's stock price. While able to model the relationships shown, the CEV model would need a time-varying β, so this model is also not appropriate. The extended compound model combines the compound and displaced diffusion models' assumptions, and through changes to the firm's debt structure and its asset mix, is able to motivate a changing leverage effect through

Year	B_t	S_t	LR_t	Elasticity
1993	1336.4	6542.73	0.204	+
1994	1563.3	9655.85	0.162	0
1995	1470.4	11507.66	0.128	0
1996	1297.1	11715.52	0.111	+
1997	1809.6	12395.77	0.146	+
1998	2038.4	15713.07	0.130	−
1999	2251.0	15978.45	0.141	0
2000	4323.0	15161.73	0.285	−
2001	5481.0	11294.35	0.485	+

Table 4.3. Debt and leverage ratios for Telecom New Zealand. The level of debt B_t, measured in millions of NZ$, is obtained from Datastream (2002) records and is "Total debt" under their classification scheme. It combines short-term and long-term debt. The value of equity S_t, also measured in millions of NZ$, is obtained from Datastream records and is "Market value". It is found using the number of outstanding shares times the share price on the balance sheet date. The leverage ratio is defined in (4.1). Balance sheet dates are typically at 31 March of the stated year, except for 2000 and 2001, taken at 30 June. The elasticity is based on the estimated slopes in Figure 4.13, and are positive (+), negative (−) or insignificant (0).

time. With reference to Telecom's financial statements, and the relationships seen in Figure 4.13 we explore the suitability of this model.

Leverage ratios are estimated for Telecom for each of the calendar years in the 1993-2001 period. This is done fairly crudely, and the results are summarised in Table 4.3. The debt figures B_t are the *Total debt* as defined by Datastream (2002), and these are collected from Telecom's balance sheet statements in the respective annual reports. Total debt figures for 1993-99 are from the 31 March annual reports, and for 2000-01 are from annual reports to 30 June. The equity figures S_t are also provided by Datastream and are calculated on the date of the financial statement by multiplying the number of outstanding shares by the market share price. The leverage ratio is calculated using (4.1), and we find the estimated leverage ratios decrease monotonically from 1993 to 1996, and then increase over the remaining years. A high ratio in 1997 prevents the increase being monotonic; however the trend is nonetheless very clear.

Spearman's rank order correlation coefficient suggests a very weak relationship between the estimated leverage ratios LR_t in Table 4.3 and the sign of the leverage relationships estimated in Figure 4.13 using the volatility estimate; however the agreement between these two quantities is loosely consistent with the extended compound model. Under this model, a high leverage ratio would suggest that the risk-free assets of the firm were insufficient to cover all debt, and hence a negative

relationship between log volatility and log share price. Conversely, a low leverage ratio would increase the chance of risk-free debt and hence an increasing relationship in Figure 4.12. Of course, the leverage figures given in Table 4.3 are just snapshots of the firm at a specific date, and so do not reflect the structure of the firm over the whole calendar year. Nonetheless, we see that where significant elasticity occurs, 50% of the time, the LR_t estimates are as we would expect: in 1996 and 1997, leverage is low and we see a positive elasticity, and in 2000, leverage is high and we see a negative elasticity. In 1993, the firm has a small absolute debt level, but also S_t is small inducing a large leverage ratio, inconsistent with the positive elasticity. In 1998, the debt level increases and the leverage relationship implies this debt is risky. However, a sharp increase in equity value causes the leverage ratio to decrease. The very large leverage ratio in 2001 is induced by low equity value, but also a large level of debt. This particular year is least consistent with the model we are proposing.

We briefly focus on the years 1997 and 2000. The share price data, volatility and leverage plot are shown for the 1997 period in Figure 4.14. Superimposed on the leverage plot are the regression line, and the robust non-parametric relationship estimated by `loess`. Both of these procedures suggest an increasing relationship between volatility and price, and the general agreement between the two supports use of a CEV model with $\beta > 2$ for this period. From Table 4.3 we see that in 1997, Telecom had a relatively low leverage ratio, and this helps explain the positive relationship between volatility and stock price, suggesting excess risk-free assets, and suitability of the displaced diffusion model (nested in the extended compound model, with $k = 0$).

In contrast to 1997, 2000 was a period in which Telecom had a much higher leverage ratio, both due to much higher debt levels, and also to lower equity value. The price series, volatility and leverage plots for 2000 are shown in Figure 4.15, and a clearly negative relationship between volatility and stock price is apparent. Once again, the relationship is approximately linear; however this time the CEV parameter would be consistent with the classical leverage effect. The slope of the regression line in the leverage plot of Figure 4.15 is -0.51, yielding an estimate of $\beta = 2 \times \text{slope} + 2 = 0.98$. Also consistent with the volatility-price relationship is the extended compound model with $k > 0$ which models a firm with risky debt and heterogeneous assets.

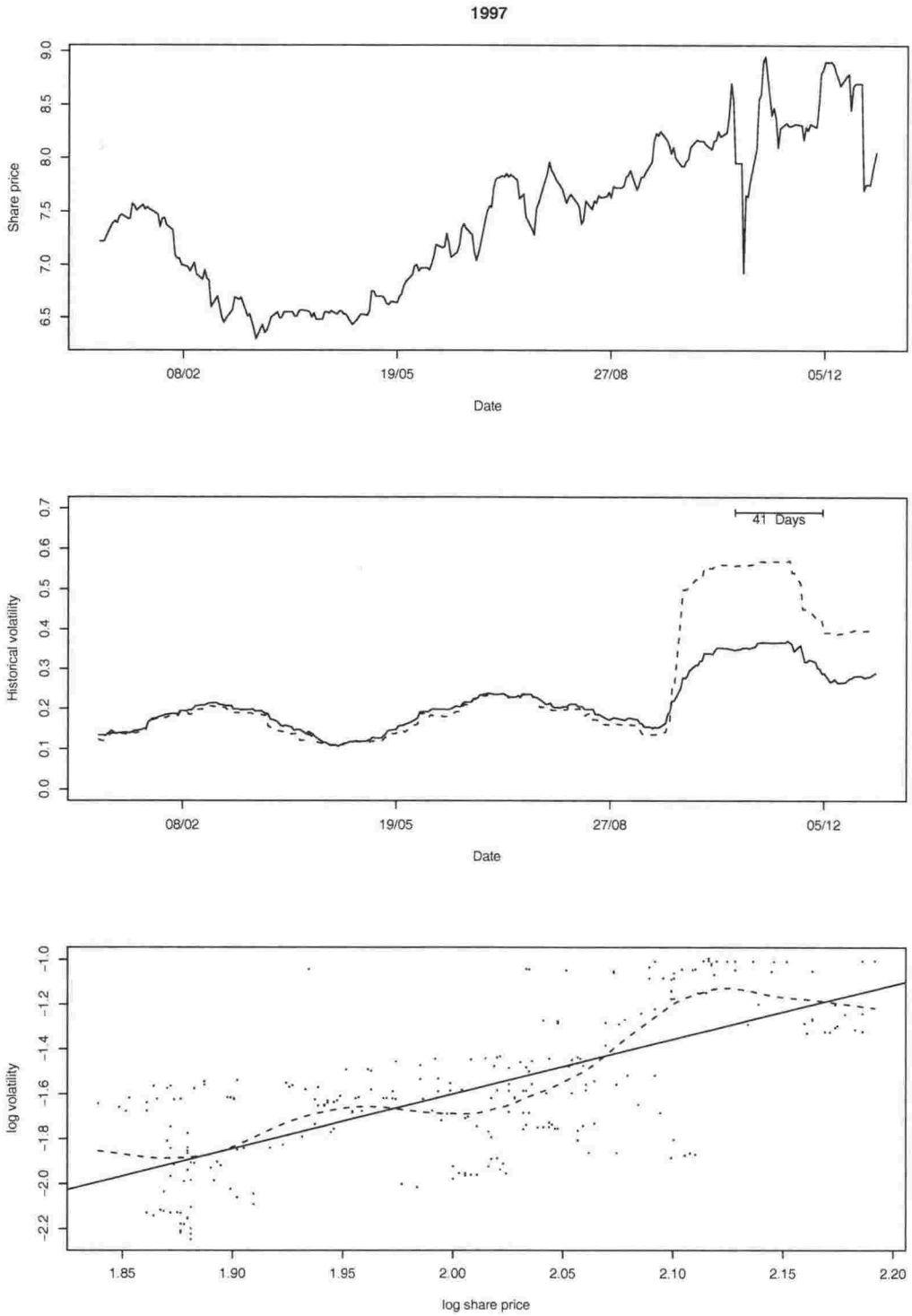

Figure 4.14. Telecom Corporation of New Zealand share price, volatility, and elasticity relationship for the year 1997. The top plot shows the share price series. The second plot gives the volatility estimates obtained using the iterated t-volatility estimator (using the solid line) and for the moving standard deviation (using the dashed line) for a window width of 41 observations. The third plot shows the log of iterated t volatility estimate against log share price, the estimated linear relationship between them, and the non-parametric relationship estimated using `loess` with a smoothing window of $\frac{1}{3}$.

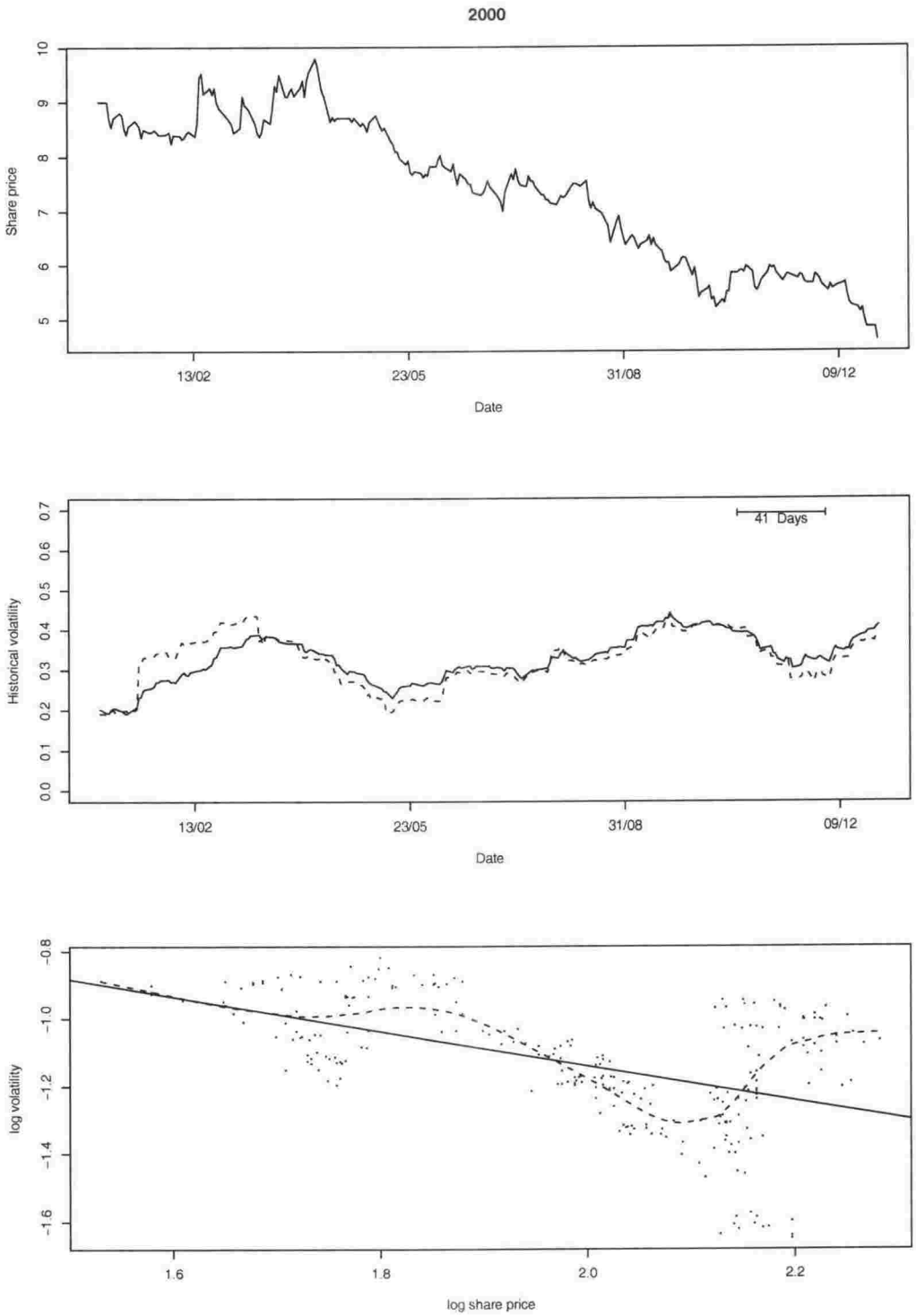

Figure 4.15. Telecom Corporation of New Zealand share price, volatility, and elasticity relationship for the year 2000. The top plot shows the share price series. The second plot gives the volatility estimates obtained using the iterated t-volatility estimator (using the solid line) and for the moving standard deviation (using the dashed line) for a window width of 41 observations. The third plot shows the log of iterated t volatility estimate against log share price, the estimated linear relationship between them, and the non-parametric relationship estimated using `loess` with a smoothing window of $\frac{1}{3}$.

4.4 Conclusions

We have successfully amalgamated the structural assumptions of Geske (1979) and Rubinstein (1983) with extension to allow multiple debt payments. The resulting compound option pricing formula enables us not only to price call options for firms with heterogeneous assets and risky debt, but also allows us to price the outstanding debt. On the basis of the results in Table 4.2, we see that the extended compound option pricing model has flexibility to address the leverage effect identified in stock price volatilities and embodied in the volatility smile. In order to achieve this flexibility, the model has a greater number of parameters than other closed-form option pricing models; however debt structure may be relatively straightforward to approximate using the firm's financial statements. As discussed by Rubinstein, α_t and σ may be best inferred from price data. He suggests time series data of S_t; however an alternative is to imply the parameters using market stock and call prices. As pointed out by Rubinstein, once α_t is estimated, its path through time follows directly from the stock price path.

Rubinstein (1983) commented on the implausibility of decomposing the assets of a firm into a single class of homogeneous risky assets and non-risky assets; however we feel that while unrealistic, this particular decomposition more closely reflects reality than competing models. Attempts to relax this assumption to allow two correlated classes of risky assets (low risk and high risk perhaps) fail to produce a simple formula for call price.

Analysis of historical data for Telecom Corporation of New Zealand Ltd shows some empirical support for the model. Both positive and negative elasticity relationships are found in subsamples of the data, and these shifts are qualitatively partly explained by crude estimates of the leverage ratio over these periods. Given the inability of any other option pricing model to successfully explain such a changing relationship through time (the CEV model shows promise for short-term modelling of the stock price; however, the elasticity parameter β is time-varying), we feel that the extended compound model makes a significant contribution, but it needs to be subjected to a more rigorous empirical treatment. In particular, we need to see whether the model can resolve the volatility smile, and other systematic biases found in the Black-Scholes option prices.

Chapter 5

Summary

Bearing in mind the original motivation of this thesis, which was estimation of the leverage effect based on a robust volatility estimator, we have accomplished more than we set out to. The highlights of this thesis for the author are the iterated t-volatility estimator of Chapter 3, and Theorem 4.7, in which a new closed form European call-option pricing model is derived.

The thesis begins by briefly demonstrating the difficulties associated with using the robust smoother `loess` to estimate time-varying volatility of a price series. Many favoured robust techniques use robustness weights to downweight extreme observations. As shown, this works well when estimating location on the basis of a symmetric distribution; however this method of downweighting is not appropriate when the distribution in question is highly skewed, e.g. when estimating $\mathrm{E}(R_t^2)$, where R_t is the daily return. This example leads us to consider specialized techniques for robustly estimating the variability, or scale, of data.

Using the limited computing power available in the early 1980s, Lax (1985) undertook a simulation study to estimate the finite sample efficiency of robust scale estimators. His results showed that the biweight A-estimator of scale, which uses the median absolute deviation (MAD) as an auxiliary robust scale estimator, is highly efficient for samples from three "extreme" situations: the normal, one-wild and slash. More significant to Lax's results than the relatively poor computing available at the time, was the evaluation criterion used: Tukey's triefficiency. Because triefficiency is based on the minimum efficiency over the three corner distributions, it is critical that the individual efficiencies are computed relative to the minimum variance estimator. In the case of the normal distribution, this minimum variance estimator is just the sample standard deviation; however, in the one-wild and slash situations, the minimum variance estimators were not used by Lax.

In Chapter 2, we derive recursions for the maximum likelihood (ML) scale estimator for a one-wild sample, and confirm known results for the slash distribution. These recursions are implemented using the EM algorithm and applied to simulated data. The remaining estimators are then benchmarked against the sampling variability of these ML estimators for each of the three corners, allowing the triefficiencies to be correctly calculated.

In addition to the ML estimates for the normal, one-wild and slash situations, we derive the ML scale estimator for the family of t-distributions. We investigate use of this estimator with a prespecified degrees of freedom parameter ν, as a general purpose scale estimator. Two forms of this estimator are considered: a fully iterated (ML) estimator, and a one-step estimator based on an auxiliary robust scale estimator, and one iteration of the EM algorithm. These estimators will be particularly suitable if the true distribution of the data is close to the specified t_ν distribution; however we hope that this distribution will be a reasonable compromise distribution for the three corners, and the estimators based on it useful more generally.

A large scale simulation study is conducted and reported in Chapter 2, in which samples are randomly generated from each of the corner distributions, various scale estimates computed for each of these samples, and then the sampling variability of these estimates compared. Triefficiencies are computed for each of the estimators considered, and these are used to evaluate the quality of each of the estimators. We find that, using the ML estimates as the benchmark, triefficiencies are generally lower than those reported by Lax (1985), due largely to poor performance of the estimators for one-wild samples. The conclusions of Rousseeuw & Croux (1993) are confirmed, and the statistics S_n and Q_n are found to be more efficient estimators of scale than the MAD. These estimators are useful in their own right, and also as auxiliary estimators in more complicated estimators. Overall, the best performing estimators were: the biweight A-estimator using Q_n and a scaling constant of $c = 11$, the one-step t-estimator using Q_n and a scaling constant of $c = 4.25$, and the biweight A-estimator using S_n and a scaling constant of $c = 7$. Each of these estimators had an average triefficiency in excess of 80%.

In addition to the results for the scale estimators, results for three commonly used location estimators are reported in Appendix B. These results are also influenced by the choice of minimum variance estimator, and together with the results of Chapter 2, suggest that further analysis of robust location and scale estimators be undertaken.

In Chapter 3, we move from a statistical focus to a financial focus; in particular, we investigate robust estimation of time-varying volatility. Volatility measures the standard deviation of financial returns, and thus a volatility estimator is a scale estimator. We specify a simple model for price returns, in which the returns have a smooth, time-varying volatility, an assumption which is generally consistent with the stylised facts of many financial time series.

Based on empirical regularities in returns data, we choose the t-distribution with five degrees of freedom as a candidate for the data generation process for returns, and estimate a slowly changing volatility on this basis. We form a volatility estimator based on the maximum likelihood estimator for a sample from the scaled t_5 distribution. A correction factor is developed so that the estimated innovations for the data have unit variance, allowing identification of the unobserved volatility component. This correction is based on the sample variance, and it is assumed that while inefficient over the smoothing window, the variance will be efficient for a sample the size of the entire series.

The results of Chapter 2 allow us to benchmark the iterated t-volatility estimator against a high quality robust scale estimator, in the biweight A-estimator using Q_n and a scaling constant of $c = 11$. We simulate returns with a smooth volatility function, with innovations sampled from the t-distributions with $\nu \in \{3, 5, 9\}$ and the normal distribution. Our estimator is found to perform very well in all situations, and provides estimates which indeed provide a close description of the underlying volatility for series where this function is known. We notice that the weights used to achieve robustness also act as smoothing weights, and that this improves the quality of the resulting volatility estimates.

We also apply the iterated t-volatility estimator to real price series, and note that this provides estimates similar to the historical volatility estimator (based on the moving sample standard deviation) when the returns are approximately normal. When the returns are highly leptokurtic, the iterated t-volatility estimates are less affected by the small returns often observed, and also by occasional extreme returns. While long-memory is present in the absolute return series, this is successfully accounted for by the volatility estimates, so that absolute standardised returns appear random.

Based on our simulation results, and the appearance of volatility estimates for real data, we advocate the use of the iterated t-volatility estimator with $\nu = 5$ degrees of freedom generally. We feel that the properties of this estimator are such that

quality estimates will result even if the underlying distribution of returns is not the t_5 distribution.

Chapter 4 provides analysis of four well known option pricing models: the Black-Scholes model (Black & Scholes 1973), the CEV model (Cox & Ross 1976), the compound option pricing model (Geske 1979) and the displaced diffusion model (Rubinstein 1983), the latter three models all incorporating Black-Scholes as a special case. The form and properties of the volatility functions for each of the models are discussed.

The CEV model is a "non-parametric" attempt to model leverage effects observed in financial returns, in the sense that no specification of debt is included in the model. In contrast, the compound and displaced diffusion models explicitly model debt. In particular, the compound model allows a single risky debt payment, and this results in a negative theoretical relationship between volatility and stock price. Addressing the other side of the balance sheet of the firm, the displaced diffusion model decomposes the firm's assets into risky and risk-free assets. The presence of risk-free assets allows the introduction of risk-free debt, and a positive relationship between volatility and stock price results.

We are able to combine the heterogeneous asset decomposition of the displaced diffusion model, and the risky debt assumption of the compound model to derive the extended compound option pricing model. This has applications to European call option pricing, as well as the pricing of debt and equity securities. We show that the model has the ability to explain both positive and negative relationships between stock price volatility and stock price level.

We provide a brief analysis of a single stock. Volatility is computed for this stock over a nine year period, and the observed relationship between log volatility and log price estimated for each calendar year. We see increasing, decreasing and constant relationships which seem approximately linear in many cases, lending support to the CEV model for short-term modelling of the stock price. However, the CEV model is deemed inappropriate for long-term modelling since the elasticity parameter β is strongly time-varying. The extended compound model remains a candidate model: it has the ability to model a decreasing relationship through the risky debt features of the model, a constant relationship through the Black-Scholes special case, and an increasing relationship through the displaced diffusion special case. Unlike the parameters of the CEV model, it is plausible that debt and asset mix parameters

change through time, justifying a changing leverage relationship. Certainly, the extended compound model will need to be subjected to more rigorous empirical testing.

The author hopes that this thesis will make a useful start to an academic career, and also make a useful contribution to existing statistical and financial literature.

Appendix A

The smoothing algorithm `loess`

The most general application of the smoothing algorithm `loess` is to provide a robust non-parametric estimate of the relationship between a dependent variable and p independent variables. In the following sections, we outline the technical details of the algorithm, and demonstrate its application to simulated time series data.

A.1 Analysis of the algorithm

The algorithm is implemented in the statistical software `S-PLUS` (see for example Venables & Ripley 1999) and in `R` (Ihaka & Gentleman 1996). For a single predictor variable, it is a robust scatter-plot smoother. It has been developed in stages, the first of which was Cleveland's (1979) `lowess`, which is an acronym for local weighted regression. Further development resulted in `loess`, documented in Cleveland et al. (1992). `loess` will be used in this thesis in two contexts: firstly in the general context of smoothing points (x_i, y_i), $i = 1, \ldots, n$, and secondly smoothing time series observations for which the x ordinates are equally spaced. Hence this discussion will outline the use of `loess` in the two dimensional case.

In particular, we wish to use the observations (x_i, y_i), $i = 1, \ldots, n$ to estimate the function $g(x)$ in the relationship

$$y_i = g(x_i) + \epsilon_i \tag{A.1}$$

where ϵ_i are the model innovations, assumed to be independent symmetrically distributed random variables with zero mean and constant variance σ^2. We do not parametrically specify the relationship $g(x)$, however it is assumed to be locally linear or locally quadratic in the neighbourhood of each observation x_i.

A.1.1 Non-robust smoothing

Assuming $g(x)$ is locally linear, for a fixed x_j the locally linear relationship is chosen to minimise the weighted sum of squares

$$\sum_{i=1}^{n} w_j(x_i)(y_i - \alpha_j - \beta_j x_i)^2 \tag{A.2}$$

where $w_j(x_i)$ is the neighbourhood weight (possibly zero) given to the observation y_i when the locally linear function $g(x)$ is being estimated at x_j. This is repeated for each j to provide an estimate of $g(x)$ at each of the observations, and the fitted values

$$\hat{y}_j = \hat{g}(x_j) = \hat{\alpha}_j + \hat{\beta}_j x_j.$$

Under the assumption that the ϵ_i are Gaussian, the local estimate at each point will be a weighted mean closely related to the sample mean, and hence close to optimal. The neighbourhood weights are designed to ensure that $\hat{g}(x)$ will be smooth.

Suppose we are estimating the function $g(x)$ at the point x_j. We define the distance of each observation x_i, $i = 1, \ldots, n$ from x_j as

$$d_j(x_i) = |x_i - x_j|$$

and this will be used to quantify the proximity of points in the (x, y) plane to the point of interest. The neighbourhood weights $w_j(x_i)$ for $i = 1, \ldots, n$, will be obtained from a function W, chosen so that the weights are non-negative, symmetric, and decrease as the distance $d_j(x_i)$ increases. The neighbourhood weight function used in `loess` is the triweight function

$$W(z) = \begin{cases} (1 - |z|^3)^3 & |z| \leq 1 \\ 0 & |z| > 1 \end{cases} \tag{A.3}$$

and if $d_j(x_i)$ is used as the argument, this function satisfies the three criteria above. In fact, the weights used in (A.2) are given by

$$w_j(x_i) = W\left(\frac{d_j(x_i)}{q_j}\right) \tag{A.4}$$

where q_j is the [`f`n]th largest of the $d_j(x_i)$ and `f` is chosen to ensure smoothness, but so that the relationship $g(x)$ is approximately linear over each sub-sample containing 100`f`% of the ordered data.

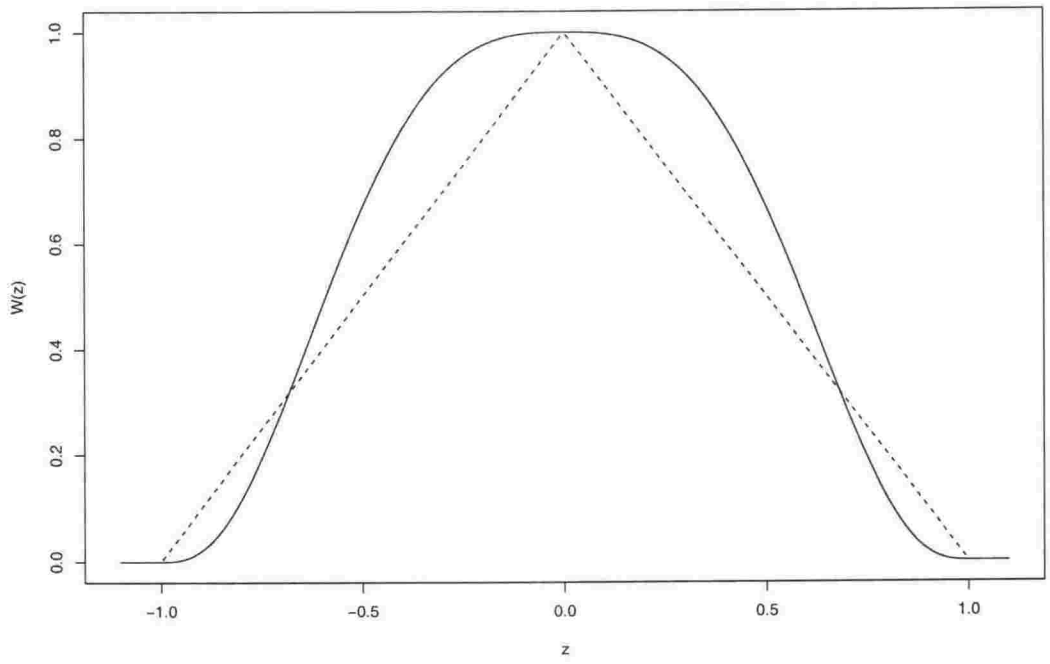

Figure A.1. The triweight function (A.3) used in `loess` to provide neighbourhood weights. Also plotted in the dashed line, is the triangular weight function $1 - |z|$ for $|z| \leq 1$.

The triweight function is shown in Figure A.1 along with the more well-known triangular function with weight $1 - |z|$ for $|z| \leq 1$, and zero otherwise. This function is also a candidate for $W(z)$, and satisfies the list of conditions above. However, use of the triweight is favoured because the resulting moving average estimate of $g(x)$ is smoother than that produced using the triangular weight function.

If no robustness properties are required in the estimate, the function $g(x)$ is estimated by (A.2), (A.4) and weighted least squares. The estimate is then a simple moving average,

$$\hat{g}(x_j) = \frac{\sum_{i=1}^{n} r_j(x_i) y_i}{\sum_{i=1}^{n} r_j(x_i)} \tag{A.5}$$

where the weights $r_j(x_i)$ depend only on the x_i, $i = 1, \ldots, n$ and x_j. This is repeated for each x_j and the relationship made continuous using linear interpolation.

One drawback of `loess` is its use of a complete smoothing window at the extremes of the data. These end-effects are discussed in the time-series context by Gray & Thomson (1990), who advocate a more traditional approach to estimation at the ends of the series, namely, reducing the length of the window, and altering weights. In contrast, at the ends of the series `loess` maintains the same "width" window, even though the observations are no longer evenly dispersed about the observation x_j at the "centre" of that window. In order to avoid issues of end-effects, where `loess`

is used in the main body of the thesis, estimates at the extremes of the independent variable (usually time) are omitted. This is particularly straightforward for time series data, where the first and last q estimates can be ignored.

An example of the application of `loess` is given in Figure A.2. For the purposes of this example, a random sample of 250 observations $X_i \sim \mathcal{N}(0, 4)$ distribution is generated, and Y_i obtained from these using

$$Y_i = \Phi(X_i) + \epsilon_i \tag{A.6}$$

where $\Phi(x)$ is the standard normal cumulative distribution function (cdf), and ϵ_i are independent normal random variables with zero mean and variance $\sigma^2 = 0.25$. The observations are plotted in Figure A.2, along with the true relationship $g(x) = \Phi(x)$, and an estimate of this given by `loess`, with non-robust smoothing. In this case, `f` is chosen to be 0.2, so for each estimate, $[\mathtt{f}n] - 1 = 49$ observations have non-zero neighbourhood weight ($[\mathtt{f}n]$ observations are less than or equal to q_j, and the observation with $d_j(x_i) = q_j$ also gets zero weight). Also plotted, in red, is a robust estimate also given by `loess`. A smoothing window is chosen so that 20% of the sample is used for each estimate of $g(x)$, and we assume that $\Phi(x)$ is locally linear over this window. From the plot, we see that the estimate provided by `loess` is very close to the true function, and also that the robust and non-robust estimates are very similar. There is some evidence of end-effects, particularly for large x_i, and in general, estimates not based on a symmetric smoothing window will not be shown when `loess` is used in this thesis.

A.1.2 Robust fitting

An alternative specification for the innovations ϵ_i is that they are identically distributed symmetric random variables, and that their distribution has heavier tails than the normal distribution. Robust estimation of $g(x)$ is sensible in this case, and is an iterative procedure that is initialised using the residuals from the standard, non-robust smooth

$$e_j = y_j - \hat{g}(x_j)$$

where $\hat{g}(x_j)$ is given in (A.5). These residuals are then used to obtain robustness weights for each observation y_j, $j = 1, \ldots, n$. Unlike the neighbourhood weights, which were functions of the distances $d_j(x_i)$, the robustness weight for each j depends only on the residual e_j and consequently is related to y_j.

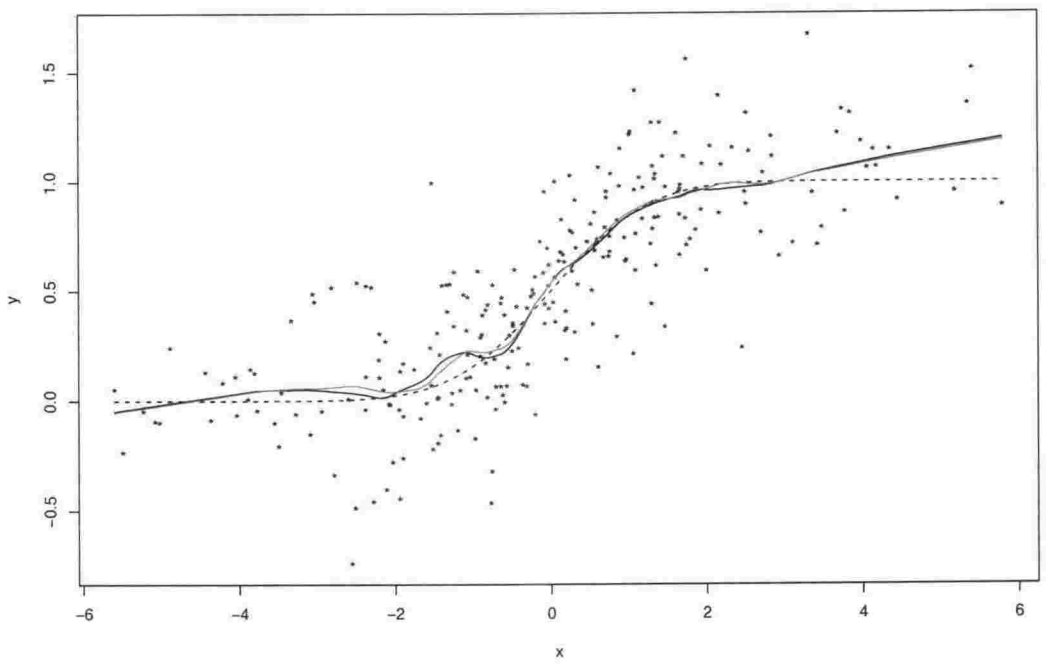

Figure A.2. Scatterplot smoothing with Gaussian innovations, where the data are generated by (A.6) with $g(x) = \Phi(x)$, the standard normal cdf, and where $\epsilon_i \sim \mathcal{N}(0, 0.25)$. $g(x)$ is shown by the dashed line, and non-robust and robust estimates from `loess` are given by the solid black and red lines respectively. The smoothing window is 20% of the data for each estimate.

In `loess`, the robustness weights are calculated using the biweight function

$$B(z) = \begin{cases} (1-z^2)^2 & |z| \leq 1 \\ 0 & |z| > 1 \end{cases} \tag{A.7}$$

and in particular, the robustness weight δ_j for a point y_j is found using

$$\delta_j = B\left(\frac{e_j}{6m}\right) \tag{A.8}$$

where $m = \text{median}_{i=1,\ldots,n} |e_j|$. A graph of the biweight function is shown in Figure A.3 and is compared to the triangular and triweight functions. We see that the down-weighting given by the biweight is intermediate.

Thus, we initialise by computing (A.5) for each j, and this is used to calculate the robustness weights δ_j. For each j, the estimate $\hat{g}(x_j)$ is then updated by minimising the sum of squares

$$\sum_{i=1}^{n} \delta_i w_j(x_i)(y_i - \alpha_j - \beta_j x_i)^2 \tag{A.9}$$

by choice of α_j and β_j. This estimation procedure is related to M-estimation, which is discussed in Chapter 2 and Appendix B. Note that the robustness weights are independent of the point of interest x_j and hence the neighbourhood weights. A

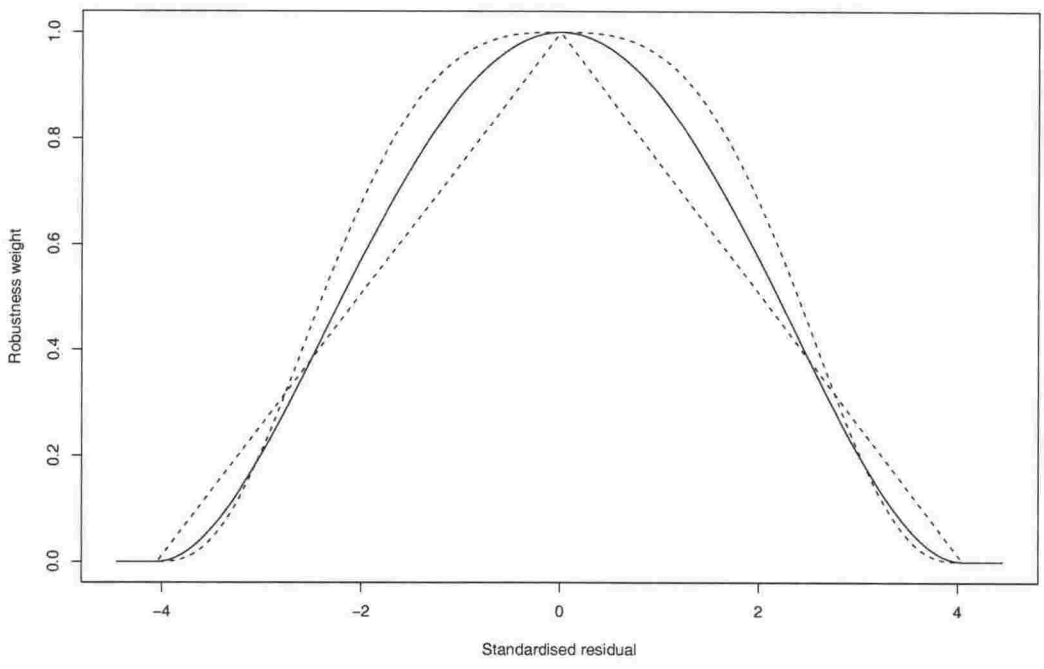

Figure A.3. The biweight function used in `loess` to provide robustness weights. Also plotted using the dashed lines, are the triangular and triweight functions.

new estimate of $g(x_j)$ results from the minimisation, and thus a new residual from which a new robustness weight will be determined.

By default, `loess` iterates this process four times. Using these robustness weights, we see that observations with large residuals have reduced influence on the sum of squares above, and hence are not as influential on the estimate $\hat{g}(x_j)$. This is very different to non-robust methods in which outlying observations which would otherwise deserve large residuals, end up having the greatest influence on the estimated model.

We continue the earlier example for symmetrically distributed heavy-tailed ϵ_i. We use the same x_i as shown in Figure A.2, and again use $g(x) = \Phi(x)$ where $\Phi(x)$ is the standard normal cdf. In this case, we contaminate the normal innovations ϵ_i, by multiplying randomly selected innovations by 10. The probability of selection is $\frac{1}{20}$, and so the ϵ_i are drawn from the contaminated normal distribution $\mathrm{CN}(\frac{1}{20}; 10)$. This distribution is discussed in Chapter 2, with an observation from it being normal with zero mean and variance σ^2 with probability $\frac{19}{20}$ and normal with zero mean and variance $100\sigma^2$ with probability $\frac{1}{20}$. In this example, 15 observations are contaminated, and six of the 250 realisations of ϵ_i lie outside $\pm 4\sigma$.

The data are once again shown in Figure A.4; however this plot is on the same scale as Figure A.2, and as a result nine of the points cannot be seen. The function $\Phi(x)$

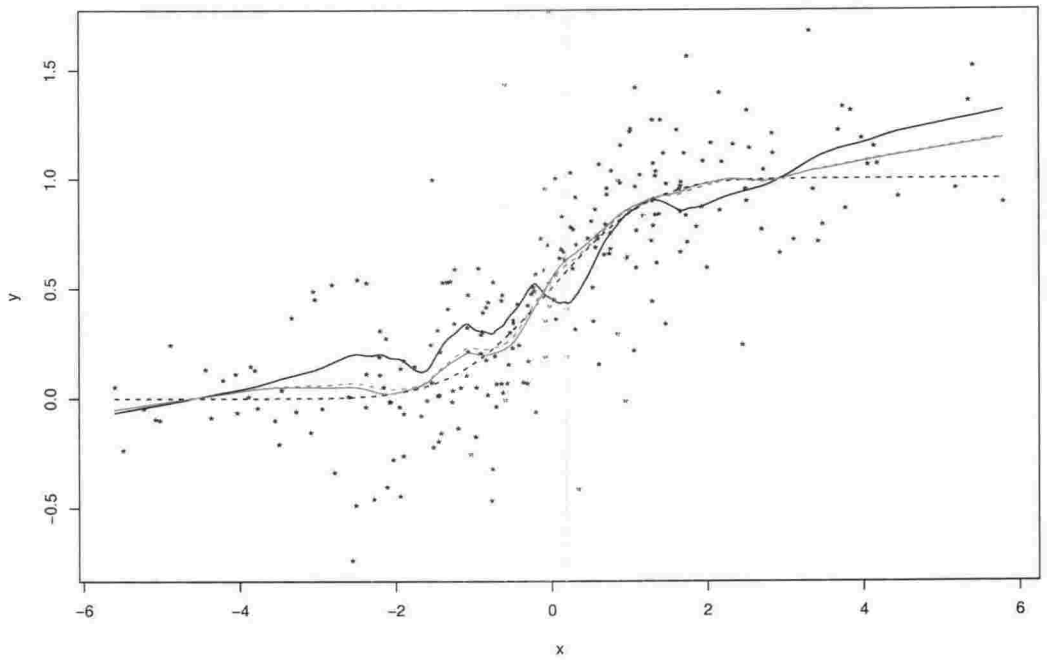

Figure A.4. Scatterplot smoothing with contaminated normal innovations, where the data are generated by (A.6) with $g(x) = \Phi(x)$, the standard normal cdf, and where $\epsilon_i \sim \text{CN}(\frac{1}{20}; 10)$, and are $\mathcal{N}(0, 0.25)$ with probability $\frac{19}{20}$, or $\mathcal{N}(0, 25)$ with probability $\frac{1}{20}$. $g(x)$ is shown by the dashed line, and the non-robust estimate from `loess` is given by the solid black line. The robust estimate is given in the solid red line, and the robust estimate from Figure A.2 is given by the dashed red line. The vertical blue lines indicate the positions of the contaminated observations. The smoothing window is 20% of the data for each estimate. Nine observations are omitted from the plot, and these extend the range to $(-1.93, 5.07)$.

is shown in the dotted line, as well as the non-robust estimate provided by `loess` in the solid black line. The vertical blue lines show the position of the outliers, and these are clearly having a large effect on the non-robust estimate. Shown in red is the robust estimate of $g(x)$ given by `loess`, and in the dotted red line, the robust relationship for the uncontaminated data. Both robust estimates are very similar, indicating the contamination has had little effect on the estimated relationship. As before, the robust estimate is a good approximation to the true function.

A.1.3 Robust smoothing for Gaussian data

It is useful to get a better feel for the robustness weights δ_j. If the data are in fact Gaussian, and we fit $\hat{g}(x)$ robustly, how different is $\hat{g}(x)$ from a non-robust fit? For Gaussian data, $\text{E}(6m) \approx 4\sigma$, and so observations greater than four standard deviations from the mean are given zero weight. Such observations occur roughly 0.006% of the time by chance, and so are very infrequently observed. In the following

theorem, we derive the probability function of a robustness weight computed using (A.8).

Theorem A.1 *The cdf of a robustness weight δ_j computed using (A.8) where $e_j = \sigma Z$ and Z is a symmetric random variable with zero mean, and probability function F, is*

$$\Pr(\delta_j < z) = \begin{cases} 0 & z < 0 \\ 2F(-\frac{6\tilde{\mu}}{\sigma}\sqrt{1-\sqrt{z}}) & 0 \leq z < 1 \\ 1 & z \geq 1. \end{cases} \tag{A.10}$$

where $\tilde{\mu}$ is the population median of $|\sigma Z|$.

Proof First, we note that since $\delta_j = B(\frac{e_j}{6\tilde{\mu}})$, it follows from (A.7) that $0 \leq \delta_j \leq 1$. Hence $\Pr(\delta_j < 0) = \Pr(\delta_j > 1) = 0$ as required.

For $0 \leq z < 1$

$$\begin{aligned} \Pr(\delta_j < z) &= \Pr\left\{\left(1 - (\tfrac{e_j}{6\tilde{\mu}})^2\right)^2 < z\right\} \\ &= \Pr\left\{\, |e_j| > 6\tilde{\mu}\sqrt{1-\sqrt{z}}\right\} \\ &= \Pr\left\{\, |Z| > \tfrac{6\tilde{\mu}}{\sigma}\sqrt{1-\sqrt{z}}\right\} \\ &= 2F(-\tfrac{6\tilde{\mu}}{\sigma}\sqrt{1-\sqrt{z}}) \end{aligned}$$

as required. □

Figure A.5 shows the cdf for the robustness weights computed using (A.8), and where Z is standard normal, and from the Student's t-distribution with $\nu = \{3, 2, 1\}$ degrees of freedom. From the graph it is clear that in the case of normal data, the majority of observations are given high weights, and in particular, only 10% of the observations are given a robustness weight less than 0.7. As ν decreases, the robustness weights become progressively smaller, and in particular, for the t-distributions, the 10th percentiles for the weights are approximately 0.55, 0.4, and 0 for $\nu = \{3, 2, 1\}$ respectively. Median weights for all distributions are approximately 0.95.

As demonstrated in Figures A.2 and A.4, there is not much difference between non-robust and robust estimates when the model innovations are Gaussian; however the differences can be quite large when the innovations are heavy-tailed. It seems that a reasonable approach would be to use the robust fitting procedure to obtain residuals which can then be examined for normality. If confidence intervals for $g(x)$

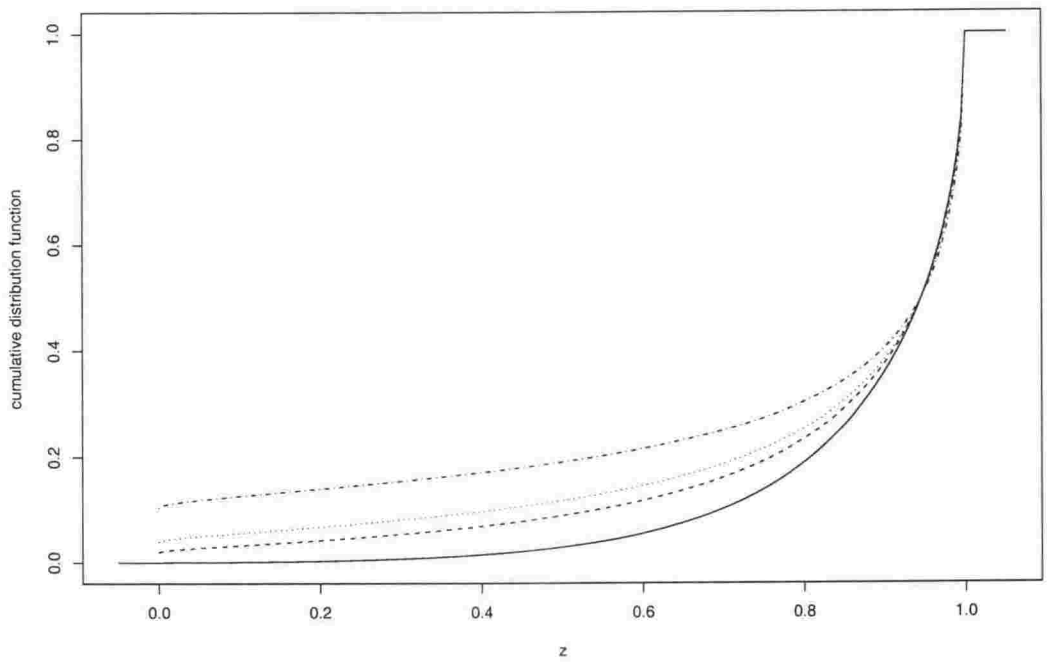

Figure A.5. The cumulative probability function of `loess` robustness weights for normal and t-distributed data. The solid line represents the cdf for standard normal residuals, the others represent t_1, t_2 and t_3 distributed residuals, from top to bottom on the left of the plot.

are sought, a decision can be made at this point whether to use the non-robust or robust forms. As far as the estimates themselves go, provided the residuals are indeed symmetrically distributed, Gaussian or not, the robust estimate of $g(x)$ will be reasonable.

A.2 Use of `loess` for time series data

In the case of time series data, `loess` performs well as a smoother, with the time series being a special case of the more general bivariate relationship. Unlike the general estimation of $g(x)$, in this case, we assume

$$Y_t = g(t) + \epsilon_t$$

where $t = 1, \ldots, T$ is the time index, and we wish to non-parametrically estimate the level of the time series $g(t)$. Since the x_i are the equally spaced sequence $t = 1, 2, \ldots, T$, the differences $x_i - x_j$ on which the smoothing weights are based are the integers $0, \pm 1, \ldots, \pm q, \pm(q+1), \ldots$, where $q_j = q = [\mathtt{f}T]$ for all j. To obtain the smoothing weights, we divide $|x_i - x_j|$ by q_j, and use this quotient in the triweight function (A.4). The differences $0, \pm 1, \ldots, \pm(q-1)$ will thus obtain a non-zero weight,

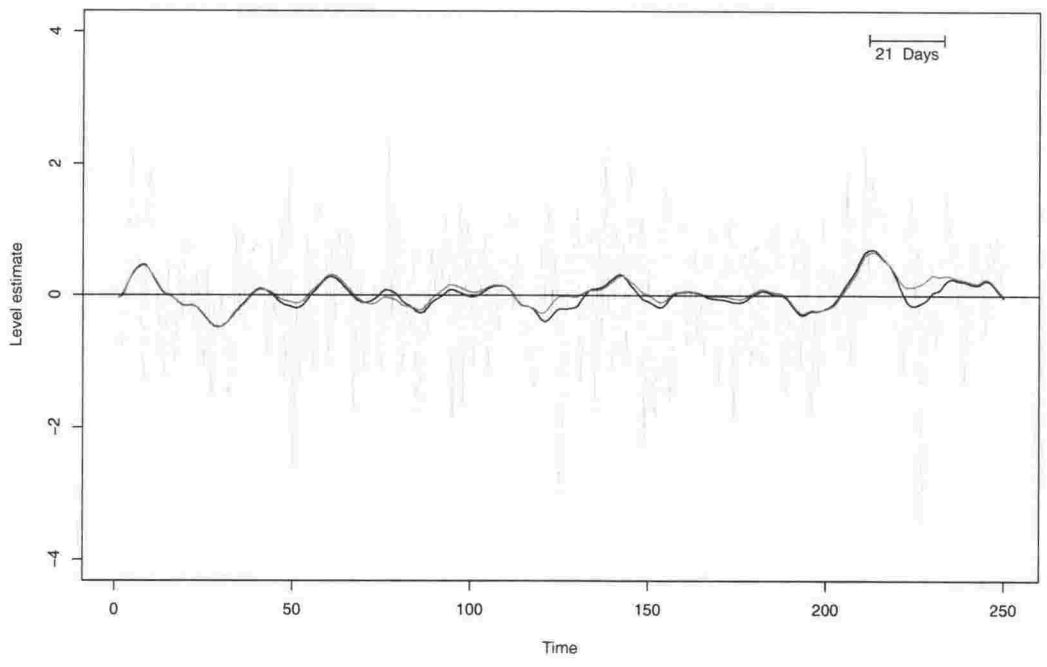

Figure A.6. Location estimates for a sequence of independent and identically distributed standard normal random variables. Two estimates are given, both computed using `loess` with a smoothing window of 21 days. The solid estimate does not have any robustness properties.

and hence the estimate of $g(x)$ is a $2q-1$ point moving average of the time series observations y_t.

We apply `loess` to a simulated Gaussian white noise series, with zero mean and constant variance $\sigma^2 = 1$. The series itself is shown in Figure A.6, along with two estimates of its level. Both the non-robust and robust estimates are very similar, and both 19-point moving averages oscillate around the true level of zero. The down-weighting of extreme observations is clearly evident in the plot, and this accounts for the differences between the two estimates. The robustness weights for the robust estimate range between 0.092 and 1.000, so none of the observations are completely omitted from the moving averages, as we would expect for Gaussian data.

We induce heavy tails in the data shown in Figure A.6 by contaminating the data. As in the earlier example, shown in Figure A.4, we multiply randomly selected observations from the Gaussian white noise process by ten. We obtain the series shown in Figure A.7, and this is a white noise process drawn from the contaminated normal distribution $\mathrm{CN}(\frac{1}{20}; 10)$, where there is a probability $\frac{19}{20}$ of an observation being unchanged, and the probability $\frac{1}{20}$ of an observation being $\mathcal{N}(0, 100)$. This new series is shown in Figure A.7, along with various estimates of its level. Since the majority of the series in Figures A.6 and A.7 are the same, it is useful to compare

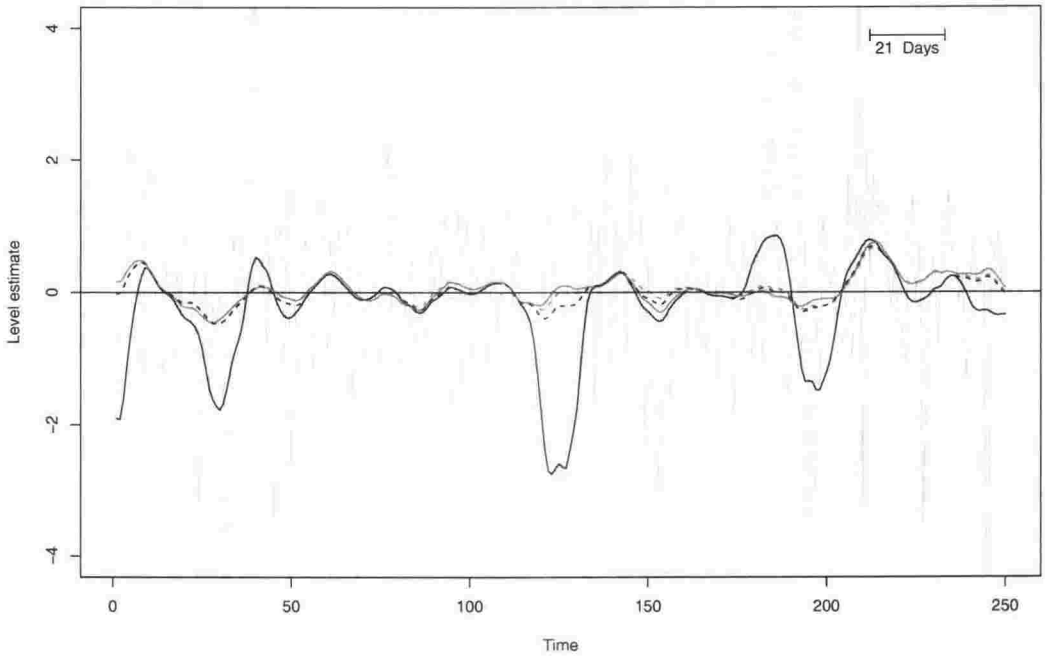

Figure A.7. Location estimates for a sequence of independent and identically distributed contaminated normal random variables, with mixing parameter $p = \frac{1}{20}$ and multiplier $k = 10$. Three estimates are given, both computed using `loess` with a smoothing window of 21 days. The non-robust estimate from `loess` is given by the solid black line. The robust estimate is given in the solid red line, and the robust estimate from Figure A.6 is given by the dashed red line.

the estimates obtained before and after contamination. In Figure A.7, the original estimates are shown by the dotted lines (in black and red for the non-robust and robust estimates respectively). In particular, we see that the non-robust estimate has been greatly affected by the outlying points. Where outlying points are present in the series, there is significant departure from the original non-robust estimate, and from the true level at zero. The two robust estimates are very similar indeed, indicating that the outlying points have not had an influence on the estimates. The new estimate is a very good approximation to the true level.

A.3 Conclusions

We conclude that `loess` provides reasonable estimates of the non-parametric relationship between two-variables, which in the special case of a time-series, is the level of the series. If the data are Gaussian, the non-robust estimate provided is a good approximation to the true underlying relationship, and in all cases, the robust estimates are a good approximation to the true underlying relationship. Thus, `loess`

should be useful for identifying a non-linear relationship in the presence of Gaussian, or heavy-tailed, symmetrically distributed errors.

Appendix B

Robust estimation of location

Chapter 2 in the main body of the thesis, motivates, describes and reports on a large simulation study of robust estimators of scale. In order to benchmark each estimator's performance in that study, we derived the form of the maximum likelihood location and scale estimators for each of the distributions considered: the normal, slash, and one-wild. The first two of these are known; however the maximum likelihood estimators of location and scale for a one-wild sample have not previously been derived.

In this appendix, we present the results of a simulation study focusing solely on location estimates. Since differences were obtained between current and previously reported results for scale estimators, this study is a simple way of checking the results for prominent location estimators as reported in Goodall (2000). All but one of the estimators considered here were used as auxiliary location estimators in the simulation study investigating scale estimators (see Chapter 2).

B.1 Location estimators

Many of the scale estimators defined in Section 2.3 feature an auxiliary estimator of location, e.g., the sample standard deviation relies on an auxiliary estimate of the sample mean; the median absolute deviation uses the sample median, and the A-estimator of scale, also dependent on the sample median, was motivated through the asymptotic variance of an M-estimator of scale. While M-estimates were not calculated during the investigation of *scale* estimators, the sample mean, median, and maximum likelihood estimates were calculated for each of the eighteen million samples simulated.

Of the location estimators we consider in this simulation, the sample mean $\bar{X}$ and the sample median M are standard. The M-estimator is formally defined as follows.

Definition B.1 (M-estimator of location) *The M-estimator of location T_n corresponding to the ψ-function $\psi(u)$, for the observations $\mathbf{X} = (X_1, \ldots, X_n)$ is the solution to the equation*

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - T_n}{cS_0}\right) = 0 \tag{B.1}$$

where $\psi(u)$ is an odd function, S_0 is an auxiliary estimate of scale, and c is a positive constant.

The sample mean and median are easy to compute for any given sample; however the M-estimate typically requires a numerical procedure to determine its value for a particular sample. As a result, it is common to consider an alternative, but related estimator, the W-estimator.

To find the W-estimator corresponding to an M-estimator, we substitute $\psi(u) = u\,w(u)$ in (B.1). Since $\psi(u)$ is an odd function, $w(u)$ is an even function, i.e. symmetric about $u = 0$. Thus, we find

$$\sum_{i=1}^{n} \left(\frac{X_i - T_n}{cS_0}\right) w\left(\frac{X_i - T_n}{cS_0}\right) = 0$$

which can be rearranged to give

$$T_n = \frac{\sum_{i=1}^{n} w\left(\frac{X_i - T_n}{cS_0}\right) X_i}{\sum_{i=1}^{n} w\left(\frac{X_i - T_n}{cS_0}\right)}.$$

The W-estimator is then found by iteration of

$$T_n^{(k+1)} = \frac{\sum_{i=1}^{n} w\left(\frac{X_i - T_n^{(k)}}{cS_0}\right) X_i}{\sum_{i=1}^{n} w\left(\frac{X_i - T_n^{(k)}}{cS_0}\right)}$$

to convergence, subject to some initial values for $T_n^{(0)}$ and a given ψ- or w-function. Note that, unlike the EM recursions, there is no guarantee that the W-estimator will converge. Assuming convergence, the W-estimate is the limit of $T_n^{(k)}$ as $k \to \infty$.

Definition B.2 (Biweight w-estimator) *The biweight w-estimator of location for the observations $\mathbf{X} = (X_1, \ldots, X_n)$ is a one-step W-estimator with the biweight w-function, and is given by*

$$T_n^{(1)} = \frac{\sum_{i=1}^{n} w(U_i) X_i}{\sum_{i=1}^{n} w(U_i)}$$

where

$$w(u) = \begin{cases} (1-u^2)^2 & |u| \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$

is the biweight function,

$$U_i = \frac{X_i - M}{cS_0}$$

is the standardised score, M is the sample median, S_0 is an auxiliary estimator of scale, and c is a positive constant.

For the purposes of this simulation, we set S_0 to be the median absolute deviation, and set $c = 6$. This estimator is a simple, yet effective, location estimator. This estimator performed well in the Princeton Robustness Study (Andrews et al. 1972) and was one of the two best performing estimators identified by Goodall (2000). It is also the basis for the robust non-parametric smoother `loess`, discussed in Appendix A. This is the third location estimator considered in this simulation.

B.2 Methodology

Samples are drawn from Tukey's three corner distributions: the normal, one-wild and slash. As in the scale estimation simulations, each run of this simulation consists of 20000 independent samples of size $n = 20$ from each distribution, and this run is repeated 100 times. As before, the one-wild and slash samples are not sampled independently from the normal samples, based on 20 independent realisations from the standard normal distribution. The one-wild sample of 20 is formed by multiplying a randomly selected observation from the normal sample by 10 (appropriate since the normal random variables are unordered, and have zero mean). The slash sample is formed by dividing each of the normal observations by an independent observation from the uniform distribution on the interval $[0, 1]$. Hence samples from the three distributions are not independent, and in particular, the normal and one-wild samples differ by only a single observation.

For each sample, the sample mean, median and the biweight w-estimate are computed, as well as the maximum likelihood location estimate for that particular situation, using a fully iterated EM algorithm and the methods described in Section 2.2.2. Efficiency is computed as

$$\text{eff}(T) = \frac{\text{sample variance of } T_1^*, \ldots, T_m^*}{\text{sample variance of } T_1, \ldots, T_m} \tag{B.2}$$

where T_i^* is the maximum likelihood estimator of location for sample i from the distribution of interest, and the T_i are the m estimates obtained from the estimator of interest.

Goodall (2000) reports estimated efficiencies for these location estimators, based on earlier simulation studies. His efficiencies are computed relative to the sample mean for the normal, the w-estimator with MAD and $c = 8.8$ for the one-wild, and the Pitman estimator for the slash. The Pitman estimator of location for a sample of n independent and identically distributed observations with location μ, unit scale, and likelihood function

$$L(\mu; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \mu)$$

is

$$\hat{\mu}(\mathbf{x}) = \frac{\int_{\mu=-\infty}^{\infty} \mu L(\mu; \mathbf{x}) d\mu}{\int_{\mu=-\infty}^{\infty} L(\mu; \mathbf{x}) d\mu}.$$

This has the minimum variance within the class of location invariant estimators of μ (Mood, Graybill & Boes 1974). Evaluation of the Pitman estimates for the slash samples has not been pursued.

B.3 Results

In the following tables, we report various summary statistics for the estimators: the sample mean, the sample median, the biweight w-estimator with $S_0 = \text{MAD}$ and $c = 6$, and the maximum likelihood estimator for the distribution in question.

Table B.1 features the average location estimates for the 12 estimator/distribution combinations. All averages are very close to zero, except for the sample mean's for the slash distribution. The observed bias in this case is several orders of magnitude greater than that for the other estimators for the one-wild and slash distributions. Curiously, all averages are negative; however this is just an artifact of the samples obtained.

Of greater interest is the precision of the estimates. Table B.2 gives the sample variance of the location estimates multiplied by the sample size, $n = 20$. In theory, the sample mean should have variance $1/n = 0.05$ in the normal case, and hence we would expect to observe $n\text{var}(\bar{X}) = 1$. As we see from the table, this corresponds closely to what is observed. We see the variances increasing as the tails become heavier for all estimators, but this is most pronounced in the case of the sample

estimator	normal	one-wild	slash
sample mean	−0.00029	−0.00072	−2.81703
sample median	−0.00042	−0.00045	−0.00065
biweight w-estimator with MAD and $c = 6$	−0.00044	−0.00043	−0.00062
maximum likelihood	−0.00029	−0.00029	−0.00046

Table B.1. Average location estimates, over 100 simulations and 20000 samples of size 20.

mean where the slash variance is over one million times greater than the next-largest variance. Also shown in the table are results from simulation studies collected by Goodall (2000) (hereafter referred to as Goodall). The results for the sample mean, median, and the Pitman estimator for the slash distribution are attributed to Andrews et al. (1972) and the biweight results are attributed to Tukey. These are based on a small scale simulation study.

Goodall's figures show near-perfect agreement for the normal samples and all three estimators as seen in Table B.2, with the greatest difference approximately −1.78% of the figure obtained for the w-estimator.

The differences for the one-wild samples are larger, and are approximately 8.9%, −4.2%, −5.0% and 2.9% for the sample mean, median, w-estimator and maximum likelihood estimator respectively, again, as a proportion of the new figures. In particular, we note that the smallest variance attained in the Goodall study is larger than the average variance of the maximum likelihood estimates. In the one-wild situation, Goodall's figure is from the biweight w-estimator with MAD and $c = 8.8$, and this is clearly not optimal for the one-wild distribution.

The discrepancies for the slash distribution are 1.6% for the median and 2.2% for the minimum variance estimator, but 15.1% for the w-estimator. It is unclear how such a large difference has arisen in this latter case, but it seems unlikely that it is due to sampling error considering the truncating weights used. Once again, we note that the smallest attained variance in the Goodall study, obtained from the Pitman estimator, is larger than the average variance of the maximum likelihood estimates, again suggesting an overstated numerator in the efficiency calculations which follow, an effect which will tend to inflate efficiency.

Efficiencies are computed by dividing the smallest available variance by the variance for the alternative estimator. The average efficiencies can be calculated directly from Table B.2 using the figures given in the maximum likelihood row for the numerator

estimator	normal		one-wild		slash	
sample mean	0.999	(1.000)	5.955	(6.485)	∞	(∞)
sample median	1.480	(1.498)	1.623	(1.555)	6.491	(6.600)
biweight using MAD and $c = 6$	1.179	(1.158)	1.239	(1.177)	5.897	(6.790)
maximum likelihood	0.999	(1.000)	1.093	(1.125)	5.432	(5.552)

Table B.2. Average variance of location estimates over 100 simulations times $n = 20$. Each variance is the sample variance of the estimates from 20000 samples of size 20, and these are averaged over the 100 trials to give the figures in the table. The figures in parentheses are from Goodall (2000). In the case of the maximum likelihood row, Goodall's figures are based on the sample mean, the biweight w-estimator with MAD and $c = 8.8$, and the Pitman estimator, for the normal, one-wild and slash distributions respectively, i.e., they are not maximum likelihood.

of (B.2) for each distribution. The efficiencies in the Goodall study are computed relative to the biweight w-estimator with MAD and $c = 8.8$, and the Pitman estimator for the one-wild and slash respectively. In the normal case, the maximum likelihood estimator is used by Goodall. Average efficiencies from this simulation are given in Table B.3, and are compared to Goodall's results shown in parentheses. In addition the efficiency distributions from the simulation are shown in Figure B.1.

The results in Table B.3 seem to suggest that under the triefficiency criterion, $c = 6$ is not the optimal constant for the biweight M-estimator. This is due to the large differences between the normal and the slash efficiencies. As with the biweight A-estimators, increasing the scaling constant c improves the efficiency for the normal distribution, but decreases it for the slash. Thus, a larger scaling constant than $c = 6$ is likely to increase the triefficiency of the M-estimator, since the normal efficiency will increase, although the maximum triefficiency will likely depend on the behaviour for the one-wild as c increases. (Note that similar comments apply to the figures stated by Goodall; however his results imply $c < 6$ should be used.) This warrants further investigation.

The differences between the variance figures in Table B.2 have fed into the efficiencies, and in some instances, opposite errors in the numerator and denominator have induced larger differences in the efficiencies. In particular, the efficiencies for the median and w-estimator in the one-wild situation have diverged between studies.

B.4 Conclusions

In summary, these new, more extensive simulation results have not shown a great deal of change from previous triefficiency results, reported in Goodall (2000). De-

estimator	normal		one-wild		slash		triefficiency	
sample mean	100.0	(100.0)	18.4	(17.3)	0.0	(0.0)	0.0	(0.0)
sample median	67.5	(66.8)	67.3	(72.3)	83.7	(84.1)	67.3	(66.8)
biweight using MAD and $c = 6$	84.8	(86.4)	88.3	(95.6)	92.1	(81.8)	84.8	(81.8)

Table B.3. Average efficiencies for the selected location estimators, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. Each efficiency is computed using (B.2). The triefficiency given is the average over the 100 simulations, rather than the minimum average. The efficiency distributions for these estimators are shown in Figure B.1. The figures in parentheses are taken from Goodall (2000).

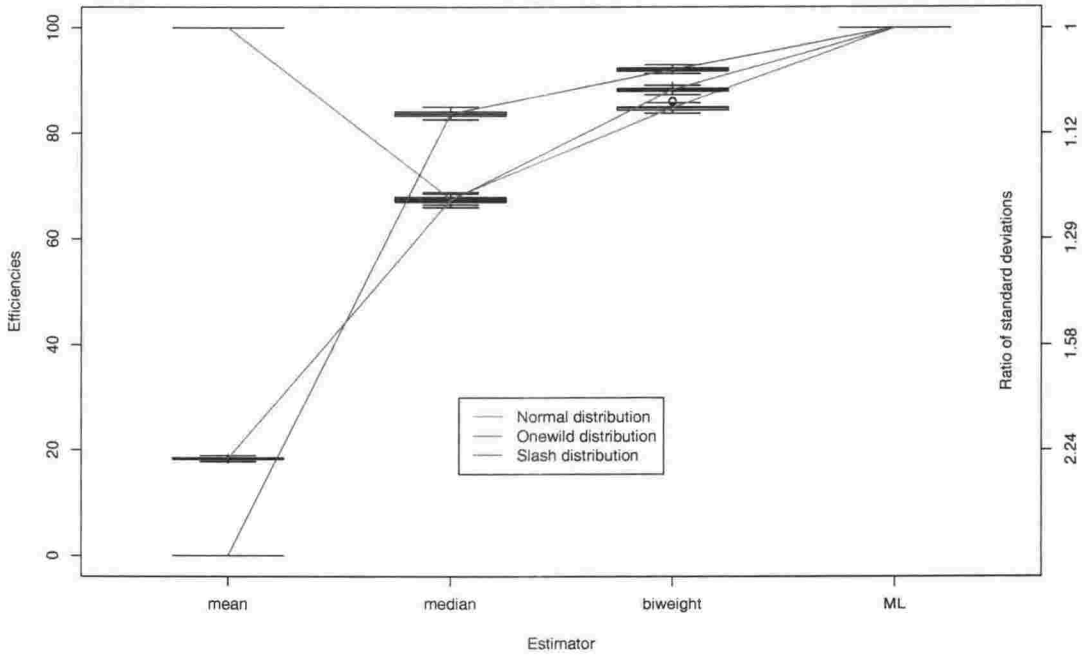

Figure B.1. Efficiency distributions for selected location estimators, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are the sample mean, sample median, biweight w-estimator using MAD and $c = 6$, and maximum likelihood. Efficiency is computed using (B.2). The ratio of standard deviations is a non-linear scale giving $1/\sqrt{\text{eff}}$, where eff is the efficiency.

spite this, efficiencies for the biweight w-estimator for the one-wild and slash distributions are quite different. This suggests $c = 6$ is not the best scaling constant to use, and a larger choice of c should result in greater triefficiency for this estimator. In each case estimators are benchmarked against maximum likelihood estimators, allowing for consistent comparison in future.

Appendix C

Scale estimation: overall results

Tables C.1, C.2 and C.3 feature the average efficiencies based on the log-variances, the average efficiencies based on the standardised variances, and the average estimates respectively, for all estimators and all distributions. The tables of efficiencies are sorted by the average rank of the respective triefficiencies over the 100 simulations, where rank is decreasing in triefficiency. Table C.1 presents the information from Tables 2.8, 2.10, 2.12, 2.13, 2.14 and 2.16 in the main body of the thesis.

estimator	normal	onewild	slash	tri	rank
maximum likelihood	100.0	100.0	100.0	100.0	1.00
biweight with Q_n and $c = 11$	89.4	82.2	82.9	82.1	3.12
biweight with Q_n and $c = 10.5$	88.0	82.1	83.9	82.1	3.62
one-step t with Q_n and $c = 4.25$	86.9	81.8	85.0	81.8	4.21
one-step t with Q_n and $c = 4$	85.7	81.7	86.2	81.7	5.19
biweight with Q_n and $c = 11.5$	90.6	82.1	82.0	81.7	5.58
one-step t with Q_n and $c = 4.5$	88.1	81.7	83.7	81.7	6.12
biweight with S_n and $c = 7$	89.0	81.1	85.8	81.1	7.55
biweight with S_n and $c = 6.5$	86.8	80.8	86.9	80.8	9.19
biweight with S_n and $c = 7.5$	90.8	80.8	84.6	80.8	9.44
biweight with MAD and $c = 10$	89.4	79.2	86.8	79.2	11.15
biweight with MAD and $c = 9$	86.2	79.1	88.0	79.1	11.84
biweight with MAD and $c = 11$	91.7	78.2	85.5	78.2	13.07
one-step t with S_n and $c = 3$	85.3	76.6	87.9	76.6	15.20
fully iterated t with $\nu = 1$	79.8	82.6	76.8	76.8	15.47
one-step t with S_n and $c = 2.75$	83.1	76.6	89.8	76.6	15.76
biweight with MAD and $c = 12$	93.4	76.5	84.0	76.5	16.11
one-step t with S_n and $c = 3.25$	87.3	76.2	86.0	76.2	17.47
modified sine with $c = 2.1$	78.1	75.3	89.0	75.3	18.91
biweight with MAD and $c = 13$	94.7	74.1	82.4	74.1	19.99
one-step t with MAD and $c = 4.25$	80.8	69.3	89.7	69.3	21.02
one-step t with MAD and $c = 4.5$	82.6	69.1	88.5	69.1	22.41
one-step t with MAD and $c = 4$	78.9	69.1	90.8	69.1	22.57
one-step t with MAD and $c = 4.75$	84.3	68.8	87.3	68.8	24.00
one-step t with MAD and $c = 5$	85.7	68.2	86.1	68.2	25.02
one-step t with MAD and $c = 5.25$	87.0	67.4	84.8	67.4	26.16
Q_n	66.9	68.4	94.9	66.9	26.84
trimmed sd with $p = r = 0.2$	65.0	70.8	76.1	65.0	28.26
fully iterated t with $\nu = 2$	85.5	86.3	64.3	64.3	28.75
trimmed sd with $p = 0.2$ and $r = 0.15$	72.1	78.6	63.4	63.4	29.97
fully iterated t with $\nu = 3$	89.0	87.1	54.9	54.9	31.45
S_n	54.7	55.9	95.8	54.7	31.55
modified biweight with $c = 6$	50.0	53.3	92.5	50.0	33.01
fully iterated t with $\nu = 4$	91.4	86.0	47.4	47.4	33.99
trimmed sd with $p = r = 0.1$	80.9	88.1	42.1	42.1	35.01
interquartile range	39.4	42.4	84.0	39.4	36.00
median absolute deviation	37.8	40.5	87.3	37.8	37.27
fully iterated t with $\nu = 6$	94.4	79.8	37.4	37.4	37.72
Gini's mean difference	98.0	26.7	11.4	11.4	39.00
sample standard deviation	100.0	11.4	7.5	7.5	40.00

Table C.1. Average efficiencies and ranks for all estimators, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are defined in Section 2.3, and each efficiency is computed using (2.31). The triefficiency given is the average over the 100 simulations, rather than the minimum average. The rank is the average rank of the 40 estimators considered, and a small rank indicates good performance.

estimator	normal	onewild	slash	triefficiency	rank
maximum likelihood	100.0	100.0	100.0	100.0	1.00
one-step t with Q_n and $c = 4$	86.2	82.0	80.9	80.9	2.62
biweight with S_n and $c = 6.5$	88.1	80.7	82.2	80.7	2.76
biweight with S_n and $c = 7$	90.1	80.6	80.8	80.4	3.98
one-step t with Q_n and $c = 4.25$	87.4	82.0	79.6	79.6	5.49
biweight with S_n and $c = 7.5$	91.6	79.9	79.5	79.3	6.41
biweight with MAD and $c = 9$	87.9	79.2	84.2	79.2	6.49
biweight with MAD and $c = 10$	90.5	78.6	82.6	78.6	8.35
one-step t with Q_n and $c = 4.5$	88.5	81.8	78.4	78.4	8.80
biweight with Q_n and $c = 10.5$	88.8	81.7	77.9	77.9	9.87
biweight with MAD and $c = 11$	92.5	77.1	81.0	77.1	12.30
modified sine with $c = 2.1$	81.2	77.0	86.3	77.0	12.57
one-step t with S_n and $c = 2.75$	83.8	76.7	85.8	76.7	13.15
biweight with Q_n and $c = 11$	90.1	81.6	76.9	76.9	13.15
one-step t with S_n and $c = 3$	85.9	76.6	83.9	76.6	14.18
biweight with Q_n and $c = 11.5$	91.2	81.2	75.9	75.9	15.95
one-step t with S_n and $c = 3.25$	87.7	76.1	82.0	76.1	16.07
biweight with MAD and $c = 12$	94.0	74.8	79.3	74.8	17.86
biweight with MAD and $c = 13$	95.1	72.1	77.7	72.1	19.00
one-step t with MAD and $c = 4$	79.9	69.4	87.4	69.4	20.30
one-step t with MAD and $c = 4.25$	81.8	69.3	86.2	69.3	20.75
one-step t with MAD and $c = 4.5$	83.5	69.0	85.0	69.0	22.08
one-step t with MAD and $c = 4.75$	85.0	68.5	83.7	68.5	23.38
Q_n	68.3	69.5	91.9	68.3	23.73
one-step t with MAD and $c = 5$	86.3	67.8	82.5	67.8	24.80
one-step t with MAD and $c = 5.25$	87.6	66.9	81.3	66.9	26.08
fully iterated t with $\nu = 1$	80.1	82.7	65.3	65.3	26.86
trimmed sd with $p = r = 0.2$	65.6	71.2	59.6	59.6	28.15
S_n	56.3	56.9	95.6	56.3	28.91
modified biweight with $c = 6$	54.0	57.1	93.6	54.0	29.95
fully iterated t with $\nu = 2$	85.6	86.3	50.1	50.1	30.99
trimmed sd with $p = 0.2$ and $r = 0.15$	72.6	78.9	44.1	44.1	32.09
interquartile range	40.6	43.5	79.0	40.6	33.08
median absolute deviation	39.2	41.9	88.6	39.2	34.41
fully iterated t with $\nu = 3$	89.1	87.2	38.9	38.9	34.41
fully iterated t with $\nu = 4$	91.4	86.1	29.6	29.6	36.00
fully iterated t with $\nu = 6$	94.3	80.4	18.9	18.9	37.27
trimmed sd with $p = r = 0.1$	81.2	88.2	17.4	17.4	37.73
Gini's mean difference	97.9	24.0	0.0	0.0	39.42
sample standard deviation	100.0	9.4	0.0	0.0	39.58

Table C.2. Average efficiencies and ranks for all estimators, based on 100 realisations of the efficiencies, each estimated from 20000 samples of size 20. The estimators are defined in Section 2.3, and each efficiency is computed using (2.32) and is based on the standardised variance. The triefficiency given is the average over the 100 simulations, rather than the minimum average. The rank is the average rank of the 40 estimators considered, and a small rank indicates good performance.

estimator	normal	onewild	slash
sample standard deviation	0.9870	2.1412	60.6327
Gini's mean difference	1.1286	1.8183	30.4786
trimmed sd with $p = r = 0.1$	0.7761	0.8434	2.9922
trimmed sd with $p = 0.2$ and $r = 0.15$	0.7083	0.7615	2.1994
trimmed sd with $p = r = 0.2$	0.6512	0.6960	1.8361
interquartile range	1.2590	1.3304	2.9667
median absolute deviation	0.6473	0.6852	1.5067
S_n	0.8582	0.9266	2.1878
Q_n	0.5360	0.5895	1.4913
modified biweight with $c = 6$	0.7843	0.8223	1.8105
modified sine with $c = 2.1$	0.9987	1.0387	2.4730
biweight with MAD and $c = 9$	0.9997	1.0510	2.6327
biweight with MAD and $c = 10$	0.9978	1.0568	2.7161
biweight with MAD and $c = 11$	0.9970	1.0646	2.7987
biweight with MAD and $c = 12$	0.9966	1.0739	2.8800
biweight with MAD and $c = 13$	0.9965	1.0846	2.9597
biweight with S_n and $c = 6.5$	1.0030	1.0521	2.6902
biweight with S_n and $c = 7$	1.0010	1.0551	2.7501
biweight with S_n and $c = 7.5$	0.9998	1.0593	2.8098
biweight with Q_n and $c = 10.5$	1.0058	1.0557	2.7872
biweight with Q_n and $c = 11$	1.0044	1.0576	2.8279
biweight with Q_n and $c = 11.5$	1.0032	1.0599	2.8684
fully iterated t with $\nu = 1$	1.3263	1.4891	4.2509
fully iterated t with $\nu = 2$	1.1364	1.3029	4.0207
fully iterated t with $\nu = 3$	1.0703	1.2527	4.1338
fully iterated t with $\nu = 4$	1.0379	1.2401	4.3444
fully iterated t with $\nu = 6$	1.0071	1.2544	4.8534
one-step t with MAD and $c = 4$	0.8241	0.9638	2.5304
one-step t with MAD and $c = 4.25$	0.8358	0.9871	2.6167
one-step t with MAD and $c = 4.5$	0.8463	1.0093	2.7001
one-step t with MAD and $c = 4.75$	0.8557	1.0306	2.7807
one-step t with MAD and $c = 5$	0.8642	1.0510	2.8588
one-step t with MAD and $c = 5.25$	0.8718	1.0706	2.9345
one-step t with S_n and $c = 2.75$	0.8100	0.9408	2.5474
one-step t with S_n and $c = 3$	0.8281	0.9739	2.6729
one-step t with S_n and $c = 3.25$	0.8434	1.0048	2.7922
one-step t with Q_n and $c = 4$	0.7910	0.9145	2.5507
one-step t with Q_n and $c = 4.25$	0.8048	0.9374	2.6382
one-step t with Q_n and $c = 4.5$	0.8172	0.9591	2.7227
maximum likelihood	0.9870	0.9606	0.9960

Table C.3. Average scale estimates, based on two million realisations of each, for samples of size 20. The estimators are defined in Section 2.3.

Appendix D

Leverage model proofs

D.1 Properties of volatility and elasticity under the compound option pricing model

In this appendix, we prove Theorems 4.2, 4.3 and 4.5. In order to facilitate these proofs, we introduce and prove additional Lemmas.

Theorem 4.2 *The stock price volatility under the compound option pricing model, $\sigma_S(V_t, t)$, defined in (4.9), has the following properties:*

1. $\sigma_S(V_t, t) > \sigma$;
2. *As* $V_t \to \infty$, $\sigma_S(V_t, t) \to \sigma$;
3. *As* $V_t \to 0$, $\sigma_S(V_t, t) \to \infty$.

Proof To prove the first property, we substitute for S_t and $\frac{\partial S_t}{\partial V_t}$ in (4.9) using (4.7) and (4.10) respectively, to give

$$\sigma_S(V_t, t) = \sigma \frac{V_t \Phi(g_t)}{V_t \Phi(g_t) - M e^{-r\tau_d} \Phi(g_t - \sigma\sqrt{\tau_d})} > \sigma$$

since $Me^{-r\tau_d} > 0$ and $\Phi(g_t - \sigma\sqrt{\tau_d}) > 0$.

To prove the second property, we note that as $V_t \to \infty$, so too does S_t, and in particular both probabilities $\Phi(g_t)$ and $\Phi(g_t - \sigma\sqrt{\tau_d}) \to 1$. Thus

$$\lim_{V_t \to \infty} \sigma_S(V_t, t) = \lim_{V_t \to \infty} \sigma \frac{V_t}{V_t - M e^{-r\tau_d}} = \sigma.$$

For the final property, L'Hôpital's Rule applies since both the numerator and denominator of $\sigma_S(V_t, t)$ go to zero as $V_t \to 0$, and the limit is given by

$$\lim_{V_t \to 0} \sigma_S(V_t, t) = \lim_{V_t \to 0} \sigma \frac{V_t}{S_t} \frac{\partial S_t}{\partial V_t} = \lim_{V_t \to 0} \sigma \frac{V_t \frac{\partial^2 S_t}{\partial V_t^2} + \frac{\partial S_t}{\partial V_t}}{\frac{\partial S_t}{\partial V_t}}.$$

Evaluating the second partial derivative, we have

$$\frac{\partial^2 S_t}{\partial {V_t}^2} = \frac{\phi(g_t)}{V_t \sigma \sqrt{\tau_d}}$$

where $\phi(x)$ is the standard normal probability density function. Now since $V_t \frac{\partial^2 S_t}{\partial {V_t}^2}$ and $\frac{\partial S_t}{\partial V_t}$ both tend to 0 as $V_t \to 0$, we apply L'Hôpital's Rule again, to find

$$\lim_{V_t \to 0} \sigma_S(V_t, t) = \lim_{V_t \to 0} \sigma \frac{V_t \frac{\partial^2 S_t}{\partial {V_t}^2} + \frac{\partial S_t}{\partial V_t}}{\frac{\partial S_t}{\partial V_t}} = \lim_{V_t \to 0} \sigma \frac{V_t \frac{\partial^3 S_t}{\partial V_t^3} + 2\frac{\partial^2 S_t}{\partial {V_t}^2}}{\frac{\partial^2 S_t}{\partial {V_t}^2}}.$$

In particular,

$$\frac{\partial^3 S_t}{\partial V_t^3} = \frac{-1}{V_t^2 \sigma \sqrt{\tau_d}} \left[1 + \frac{g_t}{\sigma \sqrt{\tau_d}} \right] \phi(g_t)$$

and hence

$$\frac{V_t \frac{\partial^3 S_t}{\partial V_t^3} + 2\frac{\partial^2 S_t}{\partial {V_t}^2}}{\frac{\partial^2 S_t}{\partial {V_t}^2}} = -\left(1 + \frac{g_t}{\sigma \sqrt{\tau_d}} \right) + 2.$$

Since $g_t \to -\infty$ as $V_t \to 0$,

$$\lim_{V_t \to 0} \sigma_S(V_t, t) = \infty$$

as required. □

Before seeking to prove Theorem 4.3, we prove Lemmas D.1 and D.2.

Lemma D.1

$$\frac{M e^{-r\tau_d} \phi(g_t - \sigma\sqrt{\tau_d})}{V_t \phi(g_t)} = 1$$

where g_t is given in (4.8), and $\phi(x)$ is the standard normal probability density function.

Proof From the definition of the standard normal density function

$$\phi(x+y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+y)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2+y^2)} e^{-xy}$$

and so

$$\frac{\phi(x-y)}{\phi(x+y)} = e^{2xy}.$$

Using the definition of g_t given in (4.8), we choose

$$x = \frac{\ln V_t - \ln M'}{\sigma\sqrt{\tau_d}} \qquad \text{and} \qquad y = \frac{1}{2}\sigma\sqrt{\tau_d} \qquad \text{giving} \qquad 2xy = \ln V_t - \ln M'$$

where $M' = Me^{-r\tau_d}$, and note that $g_t = x + y$ and $g_t - \sigma\sqrt{\tau_d} = x - y$. Hence

$$\frac{M' \phi(g_t - \sigma\sqrt{\tau_d})}{V_t \phi(g_t)} = \frac{M'}{V_t} e^{\ln V_t - \ln M'} = 1$$

as required. □

Lemma D.2 *The following limits apply:*

$$(i) \quad \lim_{x\to-\infty} \frac{\Phi(x)}{\phi(x)} = 0 \qquad (ii) \quad \lim_{x\to-\infty} \frac{\phi(x)}{x\Phi(x)} = -1 \qquad (iii) \quad \lim_{V_t\to 0} \frac{g_t S_t}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)} = -1$$

where $\phi(x)$ and $\Phi(x)$ are the standard normal pdf and cdf respectively, S_t is given in (4.7), and g_t is given in (4.8).

Proof To prove the first case, we note that L'Hôpital's Rule applies, since both numerator and denominator converge to zero, and hence

$$\lim_{x\to-\infty} \frac{\Phi(x)}{\phi(x)} = \lim_{x\to-\infty} \frac{\phi(x)}{-x\phi(x)} = \lim_{x\to-\infty} \frac{1}{-x} = 0$$

as required.

Rewriting the function of interest in the second case, and applying L'Hôpital's Rule gives

$$\lim_{x\to-\infty} \frac{\phi(x)}{x\Phi(x)} = \lim_{x\to-\infty} \frac{\frac{1}{x}\phi(x)}{\Phi(x)} = \lim_{x\to-\infty} \frac{-\frac{1}{x^2}\phi(x) + \frac{1}{x}(-x)\phi(x)}{\phi(x)} = \lim_{x\to-\infty} \frac{-\frac{1}{x^2} - 1}{1} = -1$$

as required.

The final limit is again obtained using L'Hôpital's Rule

$$\begin{aligned}
\lim_{V_t\to 0} \frac{g_t S_t}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)} &= \lim_{V_t\to 0} \frac{S_t}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)\frac{1}{g_t}} \\
&= \lim_{V_t\to 0} \frac{\Phi(g_t)}{\sigma\sqrt{\tau_d}\Phi(g_t)\frac{1}{g_t} + V_t\sigma\sqrt{\tau_d}\frac{1}{V_t\sigma\sqrt{\tau_d}}\phi(g_t)\frac{1}{g_t} - V_t\sigma\sqrt{\tau_d}\frac{\Phi(g_t)}{V_t g_t^2\sigma\sqrt{\tau_d}}} \\
&= \lim_{V_t\to 0} \frac{1}{\frac{\sigma\sqrt{\tau_d}}{g_t} + \frac{\phi(g_t)}{g_t\Phi(g_t)} - \frac{1}{g_t^2}} \\
&= \frac{1}{0 + (-1) - 0} = -1
\end{aligned}$$

by the second result of this Lemma, and since when $V_t \to 0$, $g_t \to -\infty$. □

Theorem 4.3 *The elasticity of stock price volatility under the compound option pricing model, $\theta_S(V_t, t)$, defined in (4.14), has the following properties:*

1. $\theta_S(V_t, t) > \max[-1, -(Me^{-r\tau_d})/V_t]$;
2. *As* $V_t \to \infty$, $\theta_S(V_t, t) \to 0$;
3. *As* $V_t \to 0$, $\theta_S(V_t, t) \to 0$.

Proof Since all terms in (4.13) are positive, it is clear that $\theta_S(V_t, t) > -1$. In addition, from (4.14)

$$\theta_S = \frac{-M'\Phi(g_t - \sigma\sqrt{\tau_d})}{V_t\Phi(g_t)} + \frac{S_t\phi(g_t)}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)^2} > \frac{-M'\Phi(g_t - \sigma\sqrt{\tau_d})}{V_t\Phi(g_t)} > -\frac{M'}{V_t}$$

where $M' = Me^{-r\tau_d}$, and since all terms in the second term of (4.14) are positive, and $\Phi(g_t) > \Phi(g_t - \sigma\sqrt{\tau_d})$. The maximum of these two lower bounds applies for all V_t, as required.

For the second case, we note that

$$\lim_{V_t\to\infty} \frac{S(V_t, t)}{V_t} = 1$$

since as firm value grows without bound, the present value of debt becomes a negligible proportion. In addition, as $V_t \to \infty$, $g_t \to \infty$ and

$$\frac{\partial S_t}{\partial V_t} = \Phi(g_t) \to 1 \qquad \text{and} \qquad S(V_t, t)\frac{\partial^2 S_t}{\partial V_t^2} = \frac{S(V_t, t)}{V_t}\phi(g_t) \to 1 \times 0 = 0$$

and thus from (4.13)

$$\lim_{V_t\to\infty} \theta_S = \lim_{V_t\to\infty} \left[\frac{S(V_t, t)}{V_t\frac{\partial S_t}{\partial V_t}} + S(V_t, t)\frac{\frac{\partial^2 S_t}{\partial V_t^2}}{(\frac{\partial S_t}{\partial V_t})^2} \right] - 1 = \frac{1}{1} + \frac{0}{1^2} - 1 = 0$$

as required.

For the third case, we note that as $V_t \to 0$, the numerator and denominator of both terms in (4.14) go to zero, and so L'Hôpital's Rule applies with

$$\begin{aligned}
\lim_{V_t\to 0} \theta_S &= \lim_{V_t\to 0} \frac{-M'\Phi(g_t - \sigma\sqrt{\tau_d})}{V_t\Phi(g_t)} + \lim_{V_t\to 0} \frac{S_t\phi(g_t)}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)^2} \\
&= \lim_{V_t\to 0} \frac{-M'\frac{1}{V_t\sigma\sqrt{\tau_d}}\phi(g_t - \sigma\sqrt{\tau_d})}{\Phi(g_t) + V_t\frac{1}{V_t\sigma\sqrt{\tau_d}}\phi(g_t)} + \lim_{V_t\to 0} \frac{\Phi(g_t)\phi(g_t) + S_t(\frac{-g_t}{V_t\sigma\sqrt{\tau_d}})\phi(g_t)}{\sigma\sqrt{\tau_d}\Phi(g_t)^2 + V_t\sigma\sqrt{\tau_d}2\Phi(g_t)\frac{1}{V_t\sigma\sqrt{\tau_d}}\phi(g_t)} \\
&= \lim_{V_t\to 0} \frac{-1}{\frac{\Phi(g_t)}{M'\phi(g_t - \sigma\sqrt{\tau_d})} + \frac{V_t\phi(g_t)}{M'\phi(g_t - \sigma\sqrt{\tau_d})}} + \lim_{V_t\to 0} \frac{1 - \frac{g_t S_t}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)}}{\sigma\sqrt{\tau_d}\frac{\Phi(g_t)}{\phi(g_t)} + 2} \\
&= \frac{-1}{0+1} + \frac{1-(-1)}{0+2} = 0
\end{aligned}$$

where the final result follows from Lemmas D.1 and D.2. □

Before seeking to prove Theorem 4.5, we prove the following Lemma.

Lemma D.3 *When $V_t < M$, the following limits apply:*

$$(i) \quad \lim_{\tau_d\to 0} \frac{\Phi(g_t)}{\phi(g_t)\sqrt{\tau_d}} = -\frac{\sigma}{\ln V_t - \ln M} \qquad (ii) \quad \lim_{\tau_d\to 0} \frac{S_t\phi(g_t)}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)^2} = 1$$

where $\phi(x)$ and $\Phi(x)$ are the standard normal pdf and cdf respectively, S_t is given in (4.7), and g_t is given in (4.8).

Proof For the first case, we note that since $V_t < M$, as $\tau_d \to 0$, $g_t \to -\infty$, $\Phi(g_t) \to 0$, and $\phi(g_t) \to 0$. From the definition of g_t, it follows that

$$\frac{\partial g_t}{\partial \tau_d} = -\frac{1}{2\tau_d} \frac{\ln V_t - \ln M - (r + \frac{1}{2}\sigma^2)\tau_d}{\sigma\sqrt{\tau_d}} \tag{D.1}$$

and this diverges as $\tau_d \to 0$. Applying L'Hôpital's Rule to the limit of interest

$$\begin{aligned}\lim_{\tau_d \to 0} \frac{\Phi(g_t)}{\phi(g_t)\sqrt{\tau_d}} &= \frac{\phi(g_t)\frac{\partial g_t}{\partial \tau_d}}{\frac{1}{2\sqrt{\tau_d}}\phi(g_t) - \sqrt{\tau_d} g_t \phi(g_t)\frac{\partial g_t}{\partial \tau_d}} = \lim_{\tau_d \to 0} \frac{1}{\frac{1}{2\sqrt{\tau_d}\frac{\partial g_t}{\partial \tau_d}} - \sqrt{\tau_d} g_t} \\ &= \frac{1}{0 - \frac{\ln V_t - \ln M}{\sigma}} = -\frac{\sigma}{\ln V_t - \ln M}\end{aligned}$$

by the definition of g_t, and the observation from (D.1) that $\sqrt{\tau_d}\frac{\partial g_t}{\partial \tau_d} \to -\infty$ as $\tau_d \to 0$.

For the second case, since $V_t < M$, in addition to the limits given in the proof for the first part of this Lemma, $S_t \to 0$. Thus applying L'Hôpital's Rule, and taking $\frac{\partial S}{\partial \tau_d}$ from (4.11), we have

$$\begin{aligned}\lim_{\tau_d \to 0} \frac{S_t \phi(g_t)}{V_t \sigma \sqrt{\tau_d} \Phi(g_t)^2} &= \lim_{\tau_d \to 0} \frac{Me^{-r\tau_d}\left[\frac{\sigma\phi(g_t - \sigma\sqrt{\tau_d})}{2\sqrt{\tau_d}} + r\Phi(g_t - \sigma\sqrt{\tau_d})\right]\phi(g_t) - S_t g_t \phi(g_t)\frac{\partial g_t}{\partial \tau_d}}{V_t \frac{\sigma}{2\sqrt{\tau_d}}\Phi(g_t)^2 + V_t \sigma\sqrt{\tau_d} 2\Phi(g_t)\phi(g_t)\frac{\partial g_t}{\partial \tau_d}} \\ &= \lim_{\tau_d \to 0} \frac{1 + \frac{r\Phi(g_t - \sigma\sqrt{\tau_d})2\sqrt{\tau_d}}{\sigma\phi(g_t - \sigma\sqrt{\tau_d})} - \frac{S_t g_t \frac{\partial g_t}{\partial \tau_d} 2\sqrt{\tau_d}}{\sigma V_t \phi(g_t)}}{\frac{\Phi(g_t)^2}{V_t \phi(g_t)^2} + \frac{4\tau_d \Phi(g_t)\frac{\partial g_t}{\partial \tau_d}}{\phi(g_t)}}\end{aligned} \tag{D.2}$$

where we divide by the first term in the numerator and selectively apply the identity in Lemma D.1. Since, from Lemma D.2, $\frac{\Phi(x)}{\phi(x)} \to 0$ as $x \to -\infty$, both the second term in the numerator and the first in the denominator decrease to zero at a faster rate. The third term in the numerator has the limit

$$\lim_{\tau_d \to 0} \frac{2S_t g_t \frac{\partial g_t}{\partial \tau_d}\sqrt{\tau_d}}{\sigma V_t \phi(g_t)} = \left[\lim_{\tau_d \to 0} \sqrt{\tau_d} g_t\right] \lim_{\tau_d \to 0} \frac{2S_t \frac{\partial g_t}{\partial \tau_d}}{\sigma V_t \phi(g_t)} = \frac{\ln V - \ln M}{\sigma} \lim_{\tau_d \to 0} \frac{2S_t \frac{\partial g_t}{\partial \tau_d}}{\sigma V_t \phi(g_t)} \tag{D.3}$$

from the definition of g_t and standard properties of limits. From (D.1) we note that as $\tau_d \to 0$, $\frac{\partial g_t}{\partial \tau_d} \to -\infty$ and consequently,

$$\begin{aligned}\lim_{\tau_d \to 0} \frac{2S_t \frac{\partial g_t}{\partial \tau_d}}{\sigma V_t \phi(g_t)} &= \lim_{\tau_d \to 0} \frac{2S_t}{\sigma V_t \phi(g_t)\left(1/\frac{\partial g_t}{\partial \tau_d}\right)} = \lim_{\tau_d \to 0} \frac{2\frac{V_t \sigma \phi(g_t)}{2\sqrt{\tau_d}}\left[1 + \frac{r\Phi(g_t - \sigma\sqrt{\tau_d})2\sqrt{\tau_d}}{\sigma M e^{-r\tau_d}\phi(g_t - \sigma\sqrt{\tau_d})}\right]}{\sigma V_t \left[-g_t \phi(g_t) + \phi(g_t)\frac{\partial}{\partial \tau_d}\left(1/\frac{\partial g_t}{\partial \tau_d}\right)\right]} \\ &= \lim_{\tau_d \to 0} \frac{1 + \frac{r\Phi(g_t - \sigma\sqrt{\tau_d})2\sqrt{\tau_d}}{\sigma M e^{-r\tau_d}\phi(g_t - \sigma\sqrt{\tau_d})}}{\sqrt{\tau_d}\left[-g_t + \frac{\partial}{\partial \tau_d}\left(1/\frac{\partial g_t}{\partial \tau_d}\right)\right]} = \lim_{\tau_d \to 0} \frac{1 + \frac{r\Phi(g_t - \sigma\sqrt{\tau_d})2\sqrt{\tau_d}}{\sigma M e^{-r\tau_d}\phi(g_t - \sigma\sqrt{\tau_d})}}{-\frac{\ln V_t - \ln M}{\sigma} - \frac{r + \frac{1}{2}\sigma^2}{\sigma}\tau_d + \rho(\tau_d)}\end{aligned}$$

where

$$\rho(\tau_d) = \sqrt{\tau_d}\frac{\partial}{\partial \tau_d}\left(1 \Big/ \frac{\partial g_t}{\partial \tau_d}\right)$$

is a polynomial of order τ_d with limit as $\tau_d \to 0$ of zero (from the definition of $\frac{\partial g_t}{\partial \tau_d}$ given in (D.1)). Thus

$$\lim_{\tau_d \to 0} \frac{2S_t \frac{\partial g_t}{\partial \tau_d}}{\sigma V_t \phi(g_t)} = -\frac{\sigma}{\ln V_t - \ln M}$$

and it follows from (D.3) that

$$\lim_{\tau_d \to 0} \frac{2S_t g_t \frac{\partial g_t}{\partial \tau_d}\sqrt{\tau_d}}{\sigma V_t \phi(g_t)} = -1. \tag{D.4}$$

We now turn to the final term in (D.2), and write

$$\begin{aligned}
\lim_{\tau_d \to 0} \frac{4\tau_d \Phi(g_t)\frac{\partial g_t}{\partial \tau_d}}{\phi(g_t)} &= \left[\lim_{\tau_d \to 0} 4\tau_d \sqrt{\tau_d}\frac{\partial g_t}{\partial \tau_d}\right] \lim_{\tau_d \to 0} \frac{\Phi(g_t)}{\phi(g_t)\sqrt{\tau_d}} \\
&= \lim_{\tau_d \to 0} \left[-2\frac{\ln V_t - \ln M}{\sigma} - \frac{(r + \frac{1}{2}\sigma^2)\tau_d}{\sigma}\right] \frac{-\sigma}{\ln V_t - \ln M} \\
&= -2\left[\frac{\ln V_t - \ln M}{\sigma}\right] \frac{-\sigma}{\ln V_t - \ln M} = 2
\end{aligned} \tag{D.5}$$

from (D.1) and the first result of this Lemma.

We now substitute the limits given in (D.4) and (D.5) into (D.2), and find

$$\lim_{\tau_d \to 0} \frac{S_t \phi(g_t)}{V_t \sigma \sqrt{\tau_d}\Phi(g_t)^2} = \frac{1 + 0 - (-1)}{0 + 2} = 1$$

as required. □

Theorem 4.5 *The elasticity of stock price volatility under the compound option pricing model, $\theta_S(V_t, t)$, defined in (4.14), has the following limit*

$$\lim_{\tau_d \to 0} \theta_S(V_t, t) = \begin{cases} 0 & V_t < M \\ -\frac{M}{V_t} & V_t > M. \end{cases}$$

Proof First, consider the case when $V_t > M$. As $\tau_d \to 0$, $g_t \to \infty$, where g_t is defined in (4.8). In addition $\Phi(g_t) \to 1$, and $\phi(g_t) \to 0$. Thus

$$\lim_{\tau_d \to 0} \theta_S(V_t, t) = \lim_{\tau_d \to 0} \left[\frac{-Me^{-r\tau_d}\Phi(g_t - \sigma\sqrt{\tau_d})}{V_t \Phi(g_t)} + \frac{S_t \phi(g_t)}{V_t \sigma \sqrt{\tau_d}\Phi(g_t)^2}\right] = -\frac{M}{V_t} + \frac{S_t}{V_t}\left[\lim_{\tau_d \to 0} \frac{\phi(g_t)}{\sigma\sqrt{\tau_d}}\right]$$

and since $\frac{\phi(g_t)}{\sigma\sqrt{\tau_d}}$ is the pdf of a normal random variable with zero mean and variance $\sigma^2\tau_d$ evaluated at $\ln V_t - \ln M + (r + \frac{1}{2}\sigma^2)\tau_d$, the limit as $\tau_d \to 0$ of this function is 0, and

$$\lim_{\tau_d \to 0} \theta_S(V_t, t) = -\frac{M}{V_t}$$

as required.

Second, we consider the limit when $V_t < M$. In this instance, as $\tau_d \to 0$, $g_t \to -\infty$, $\Phi(g_t) \to 0$, and $\phi(g_t) \to 0$. In addition, $S_t \to 0$. Thus

$$\lim_{\tau_d \to 0} \theta_S(V_t, t) = \lim_{\tau_d \to 0} \left[\frac{-Me^{-r\tau_d}\Phi(g_t - \sigma\sqrt{\tau_d})}{V_t\Phi(g_t)} + \frac{S_t\phi(g_t)}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)^2} \right] \tag{D.6}$$

features zero numerator and denominator in both terms, which we now consider separately. The first term is

$$\begin{aligned}
&\lim_{\tau_d \to 0} \frac{-Me^{-r\tau_d}\Phi(g_t - \sigma\sqrt{\tau_d})}{V_t\Phi(g_t)} \\
&= \lim_{\tau_d \to 0} \frac{rMe^{-r\tau_d}\Phi(g_t - \sigma\sqrt{\tau_d}) - Me^{-r\tau_d}\phi(g_t - \sigma\sqrt{\tau_d})\left(\frac{\partial g_t}{\partial \tau_d} - \frac{\sigma}{2\sqrt{\tau_d}}\right)}{V_t\phi(g_t)\frac{\partial g_t}{\partial \tau_d}} \\
&= \lim_{\tau_d \to 0} \frac{\frac{r\Phi(g_t - \sigma\sqrt{\tau_d})}{\phi(g_t - \sigma\sqrt{\tau_d})\frac{\partial g_t}{\partial \tau_d}} - 1 + \frac{\sigma}{2\sqrt{\tau_d}\frac{\partial g_t}{\partial \tau_d}}}{1}
\end{aligned}$$

where the denominator follows from Lemma D.1. As noted in the proof to Lemma D.3, as $\tau_d \to 0$, both $\frac{\partial g_t}{\partial \tau_d}$ and $\sqrt{\tau_d}\frac{\partial g_t}{\partial \tau_d}$ diverge, and from Lemma D.2 $\frac{\Phi(x)}{\phi(x)} \to 0$ as $x \to -\infty$. So the first term decreases at a faster rate, and hence

$$\lim_{\tau_d \to 0} \frac{-Me^{-r\tau_d}\Phi(g_t - \sigma\sqrt{\tau_d})}{V_t\Phi(g_t)} = -1. \tag{D.7}$$

The second term of (D.6) has limit

$$\lim_{\tau_d \to 0} \frac{S_t\phi(g_t)}{V_t\sigma\sqrt{\tau_d}\Phi(g_t)^2} = 1 \tag{D.8}$$

from Lemma D.3. Thus, substituting (D.7) and (D.8) in (D.6), we obtain

$$\lim_{\tau_d \to 0} \theta_S(V_t, t) = -1 + 1 = 0$$

confirming the second case and completing the proof. □

D.2 The compound option pricing model

In this appendix, we provide an alternative proof for Geske's (1977) compound option pricing model with generalisation to arbitrary debt repayment dates. In addition, we derive the general form of $\frac{\partial S_t}{\partial V_t}$ for this model. This latter result has application to stock price volatility estimation under this model.

Theorem D.4 *The stock pricing model of Geske (1977) can be generalised to allow general debt repayment dates, satisfying the increasing sequence $t_1 < \ldots < t_n$. Equity value at time $t < t_1$ is given by*

$$S_t = G_n(V_t, \mathbf{X}, \boldsymbol{\tau}, r, \sigma) \equiv V_t \Phi_n(h_i; \{\rho_{ij}\}) - \sum_{k=1}^{n} X_k e^{-r\tau_k} \Phi_k(h_i - \sigma\sqrt{\tau_i}; \{\rho_{ij}\}) \quad \text{(D.9)}$$

where firm value V_t is a GBM process with volatility σ, $\mathbf{X} = (X_1, \ldots, X_n)$ are fixed debt repayments with $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ where $\tau_i = t_i - t$ is the length of time until repayment i, r is the continuously compounding risk-free rate, $\Phi_n(h_i; \{\rho_{ij}\})$ is the cumulative distribution function of a standard n-variate normal random variable with correlation matrix given by $\{\rho_{ij}\}$ for $1 \leq i, j \leq n$, evaluated at $h_1, \ldots, h_n$. Also

$$h_i = \frac{\ln V_t - \ln \bar{V}_i + (r + \frac{1}{2}\sigma^2)\tau_i}{\sigma\sqrt{\tau_i}}$$

$$\bar{V}_i = \begin{cases} \text{the value of } V \text{ which solves } S_{t_i}(V) = X_i & 1 \leq i \leq n-1 \\ X_n & i = n \end{cases}$$

$$\rho_{ij} = \sqrt{\frac{\tau_i}{\tau_j}} \qquad i < j$$

and $\rho_{ji} = \rho_{ij}$.

Proof We will use an induction proof for this result. We begin by considering the case where $n = 2$, and noting that for any time t in the interval $[t_1, t_2)$, the stock price can be written

$$S_t = e^{-r(t_2 - t)} \mathrm{E}_t^{\mathbb{Q}}\{\max(V_{t_2} - X_2, 0)\}$$

where V_t is a GBM process under the risk-neutral measure $\mathbb{Q}$. It is well known that the solution to this integral equation is given by the Black-Scholes formula

$$S_t = G_1(V_t, X_2, \tau_2, r, \sigma) = V_t \Phi(h_2) - X_2 e^{-r\tau_2} \Phi(h_2 - \sigma\sqrt{\tau_2}) \quad \text{(D.10)}$$

where

$$h_2 = \frac{\ln V_t - \ln X_2 + (r + \frac{1}{2}\sigma^2)\tau_2}{\sigma\sqrt{\tau_2}}.$$

Geske (1979) shows that the stock is a compound option, with value at $t < t_1$

$$\begin{aligned} S_t &= e^{-r\tau_1} \mathrm{E}_t^{\mathbb{Q}}\{\max(S_{t_1} - X_1, 0)\} \\ &= e^{-r\tau_1} \int_{S_{t_1} > X_1} (S_{t_1}(z) - X_1)\phi(z)dz \end{aligned} \quad \text{(D.11)}$$

where S_{t_1} is a function of V_{t_1} (specifically (D.10) with $t = t_1$) which is in turn given by

$$V_{t_1} = V_t \exp\{(r - \tfrac{1}{2}\sigma^2)\tau_1 + \sigma\sqrt{\tau_1}(-Z)\} \tag{D.12}$$

where Z is a standard normal random variable. Here we choose to use $-Z$ for algebraic convenience later in the proof.

Firstly, we note that $S_{t_1} > X_1$ implies $V_{t_1} > S^{-1}(X_1) = \bar{V}_1$ where $S(V_t)$ is given by (D.10), and S^{-1} is the inverse function. Noting from (D.12) that V_{t_1} is a function of Z, we solve for the critical value of Z, and find that at $t < t_1$, $V_{t_1} > \bar{V}_1$ implies

$$Z < \frac{\ln V_t - \ln \bar{V}_1 + (r - \frac{1}{2}\sigma^2)\tau_1}{\sigma\sqrt{\tau_1}} = h_1 - \sigma\sqrt{\tau_1}.$$

We now write the explicit form for S_{t_1} using (D.10)

$$S_{t_1} = V_{t_1}\Phi(h_2') - X_2 e^{-r(\tau_2 - \tau_1)}\Phi(h_2' - \sigma\sqrt{\tau_2 - \tau_1})$$

where

$$\begin{aligned} h_2' &= \frac{\ln V_{t_1} - \ln X_2 + (r + \frac{1}{2}\sigma^2)(\tau_2 - \tau_1)}{\sigma\sqrt{\tau_2 - \tau_1}} \\ &= \frac{\ln V_t + (r - \frac{1}{2}\sigma^2)\tau_1 + \sigma\sqrt{\tau_1}(-Z) - \ln X_2 + (r + \frac{1}{2}\sigma^2)(\tau_2 - \tau_1)}{\sigma\sqrt{\tau_2 - \tau_1}} \\ &= \frac{h_2 - \sqrt{\frac{\tau_1}{\tau_2}}(Z + \sigma\sqrt{\tau_1})}{\sqrt{1 - \frac{\tau_1}{\tau_2}}} \end{aligned}$$

and also note that $e^{-r\tau_1}V_{t_1}(Z)\phi(Z) = V_t\phi(Z + \sigma\sqrt{\tau_1})$, and

$$h_2' - \sigma\sqrt{\tau_2 - \tau_1} = \frac{h_2 - \sigma\sqrt{\tau_2} - \sqrt{\frac{\tau_1}{\tau_2}}Z}{\sqrt{1 - \frac{\tau_1}{\tau_2}}}.$$

Unlike h_2', which measure time from t_1, h_2 measures time from the current time $t < t_1$.

(D.11) consists of three terms, two arising from S_{t_1} as given in (D.10) and the third relating to X_1. Utilising the above algebra, the first term of (D.11) can be written

$$\begin{aligned} e^{-r\tau_1}\int_{S_{t_1} > X_1} V_{t_1}\Phi(h_2')\phi(z)dz &= V_t\int_{-\infty}^{h_1 - \sigma\sqrt{\tau_1}} \phi(z + \sigma\sqrt{\tau_1})\Phi\left(\frac{h_2 - \sqrt{\frac{\tau_1}{\tau_2}}(z + \sigma\sqrt{\tau_1})}{\sqrt{1 - \frac{\tau_1}{\tau_2}}}\right)dz \\ &= V_t\int_{-\infty}^{h_1} \phi(z)\Phi\left(\frac{h_2 - \sqrt{\frac{\tau_1}{\tau_2}}z}{\sqrt{1 - \frac{\tau_1}{\tau_2}}}\right)dz \\ &= V_t\Phi_2\left(h_1, h_2; \sqrt{\tau_1/\tau_2}\right). \end{aligned}$$

The final equality follows from the properties of the bivariate normal cdf (Curnow & Dunnett 1962, equation 2.4), which states in general, that

$$\Phi_n(h_i;\{\rho_{ij}\}) = \int_{-\infty}^{h_1} \phi(z)\Phi_{n-1}\left(\frac{h_{i+1}-\rho_{i1}z}{\sqrt{1-\rho_{i1}^2}}\ ;\{\rho_{ij\cdot 1}\}\right)dz \tag{D.13}$$

where

$$\rho_{ij\cdot 1} = \frac{\rho_{ij}-\rho_{i1}\rho_{j1}}{\sqrt{(1-\rho_{i1}^2)(1-\rho_{j1}^2)}}$$

is the partial correlation coefficient.

The second and third terms of (D.11) become

$$\begin{aligned}\int_{-\infty}^{h_1-\sigma\sqrt{\tau_1}} &\left[X_2e^{-r\tau_2}\Phi\left(\frac{h_2-\sigma\sqrt{\tau_2}-\sqrt{\frac{\tau_1}{\tau_2}}z}{\sqrt{1-\frac{\tau_1}{\tau_2}}}\right)+X_1e^{-r\tau_1}\right]\phi(z)dz \\ &= X_2e^{-r\tau_2}\Phi_2\left(h_1-\sigma\sqrt{\tau_1},h_2-\sigma\sqrt{\tau_2};\sqrt{\frac{\tau_1}{\tau_2}}\right)+X_1e^{-r\tau_1}\Phi(h_1-\sigma\sqrt{\tau_1}).\end{aligned}$$

Thus, if there are only two outstanding debt payments, (D.11) can be written

$$\begin{aligned}S_t = V_t\Phi_2\left(h_1,h_2;\sqrt{\tfrac{\tau_1}{\tau_2}}\right) &- X_2e^{-r\tau_2}\Phi_2\left(h_1-\sigma\sqrt{\tau_1},h_2-\sigma\sqrt{\tau_2};\sqrt{\tfrac{\tau_1}{\tau_2}}\right) \\ &- X_1e^{-r\tau_1}\Phi(h_1-\sigma\sqrt{\tau_1})\end{aligned} \tag{D.14}$$

and (D.9) is thus true for $n=2$.

We now assume that the form of S_t given by (D.9) is correct for $n-1$ debt payments, and show it true for n payments. In particular, (D.9) yields the price of the stock at time $t_1 < t_2$ as

$$S_{t_1} = V_{t_1}\Phi_{n-1}(h'_{i+1};\{\rho'_{ij}\}) - \sum_{k=2}^{n} X_k e^{-r\tau'_k}\Phi_{k-1}\left(h'_{i+1}-\sigma\sqrt{\tau'_{i+1}}\ ;\{\rho'_{ij}\}\right) \tag{D.15}$$

where $\tau'_{i+1} = t_{i+1}-t_1$ for $i=1,\ldots,n-1$, V_{t_1} is given by (D.12),

$$\begin{aligned}h'_{i+1} &= \frac{\ln V_{t_1} - \ln \bar{V}_{i+1} + (r+\frac{1}{2}\sigma^2)\tau'_{i+1}}{\sigma\sqrt{\tau'_{i+1}}} \\ \bar{V}_{i+1} &= \begin{cases}\text{the value of } V \text{ which solves } S_{t_{i+1}}(V) = X_{i+1} & 1\le i\le n-2 \\ X_n & i=n-1\end{cases} \\ \rho'_{ij} &= \sqrt{\frac{t_i-t_1}{t_j-t_1}} = \sqrt{\frac{\tau_i-\tau_1}{\tau_j-\tau_1}} \qquad 1<i<j\end{aligned}$$

and $\rho'_{ji} = \rho'_{ij}$.

The value of the stock at time $t < t_1$ is given by

$$S_t = e^{-r\tau_1} \mathrm{E}_t^{\mathbb{Q}}\{\max(S_{t_1} - X_1, 0)\}$$

and substituting the form of S_{t_1} given by (D.15), we must evaluate the integral equation

$$S_t = e^{-r\tau_1} \int_{S_{t_1} > X_1} (S_{t_1}(z) - X_1)\phi(z)dz. \tag{D.16}$$

As for the case where $n = 2$, $S_{t_1} > X_1$ implies $Z < h_1 - \sigma\sqrt{\tau_1}$ and again $e^{-r\tau_1} V_{t_1}(Z)\phi(Z) = V_t\phi(Z + \sigma\sqrt{\tau_1})$. We note that with some algebra, we can write

$$h'_{i+1} = \frac{\ln V_{t_1} - \ln \bar{V}_{i+1} + (r + \frac{1}{2}\sigma^2)\tau'_{i+1}}{\sigma\sqrt{\tau'_{i+1}}} = \frac{h_{i+1} - \sqrt{\frac{\tau_1}{\tau_{i+1}}}(Z + \sigma\sqrt{\tau_1})}{\sqrt{1 - \frac{\tau_1}{\tau_{i+1}}}}$$

for $i = 1, \ldots, n-1$ and also that ρ'_{ij} is of the form

$$\rho'_{ij} = \rho_{ij \cdot 1} = \frac{\rho_{ij} - \rho_{i1}\rho_{j1}}{\sqrt{(1 - \rho_{i1}^2)(1 - \rho_{j1}^2)}}$$

for $1 < i < j$, where $\rho_{ij} = \sqrt{\frac{\tau_i}{\tau_j}}$ for $i < j$. The value of the stock at time t_1 consists of n terms (as shown in (D.15)), and an $(n+1)$th term for S_t is obtained in (D.16) through X_1. Making the appropriate substitutions in (D.16), the first term is given by

$$\begin{aligned}
&e^{-r\tau_1} \int_{S_{t_1} > X_1} V_{t_1}(z)\Phi_{n-1}(h'_{i+1}; \{\rho'_{ij}\})\phi(z)dz \\
&\quad = V_t \int_{-\infty}^{h_1 - \sigma\sqrt{\tau_1}} \phi(z + \sigma\sqrt{\tau_1})\Phi_{n-1}\left(\frac{h_{i+1} - \sqrt{\frac{\tau_1}{\tau_{i+1}}}(z + \sigma\sqrt{\tau_1})}{\sqrt{1 - \frac{\tau_1}{\tau_{i+1}}}} \; ; \{\rho'_{ij}\} \right) dz \\
&\quad = V_t \int_{-\infty}^{h_1} \phi(z)\Phi_{n-1}\left(\frac{h_{i+1} - \sqrt{\frac{\tau_1}{\tau_{i+1}}}z}{\sqrt{1 - \frac{\tau_1}{\tau_{i+1}}}} \; ; \{\rho'_{ij}\} \right) dz \\
&\quad = V_t\Phi_n(h_i; \{\rho_{ij}\})
\end{aligned} \tag{D.17}$$

by (D.13). As for the case when $n = 2$, the final term is

$$e^{-r\tau_1} \int_{S_{t_1} > X_1} X_1\phi(z)dz = e^{-r\tau_1} \int_{-\infty}^{h_1 - \sigma\sqrt{\tau_1}} X_1\phi(z)dz = X_1 e^{-r\tau_1}\Phi(h_1 - \sigma\sqrt{\tau_1}).$$

The remaining terms are given by

$$\begin{aligned}
&e^{-r\tau_1}\int_{-\infty}^{h_1-\sigma\sqrt{\tau_1}}\sum_{k=2}^{n}X_k e^{-r\tau'_k}\phi(z)\Phi_{k-1}\left(h'_{i+1}-\sigma\sqrt{\tau'_{i+1}}\ ;\{\rho'_{ij}\}\right)dz \\
&\quad=\sum_{k=2}^{n}X_k e^{-r\tau_k}\int_{-\infty}^{h_1-\sigma\sqrt{\tau_1}}\phi(z)\Phi_{k-1}\left(\frac{h_{i+1}-\sigma\sqrt{\tau_{i+1}}-\sqrt{\frac{\tau_1}{\tau_{i+1}}}z}{\sqrt{1-\frac{\tau_1}{\tau_{i+1}}}}\ ;\{\rho'_{ij}\}\right)dz \\
&\quad=\sum_{k=2}^{n}X_k e^{-r\tau_k}\Phi_k(h_i-\sigma\sqrt{\tau_i};\{\rho_{ij}\})
\end{aligned}$$

by the result (D.13), as before. Thus, combining terms we find

$$S_t = V_t\Phi_n(h_i;\{\rho_{ij}\}) - \sum_{k=1}^{n}X_k e^{-r\tau_k}\Phi_k(h_i-\sigma\sqrt{\tau_i};\{\rho_{ij}\})$$

as required, and the induction proof is complete. □

Theorem D.5 *If stock price is given by (D.9), the partial derivative with respect to firm value at time t is given by*

$$\frac{\partial S_t}{\partial V_t} = \Phi_n(h_i;\{\rho_{ij}\}) \tag{D.18}$$

where all terms and notation are defined in the statement of Theorem D.4.

Proof Once again we use an induction proof. The base case is provided by the familiar Black-Scholes hedge ratio, $\frac{\partial S_t}{\partial V_t} = \Phi(h_1)$. We now assume the theorem true for $n-1$ outstanding debt payments, and seek to prove it true for n payments.

From Theorem D.4, stock price at time $t < t_1$ is given by

$$\begin{aligned}
S_t &= V_t\Phi_n(h_i;\{\rho_{ij}\}) - \sum_{k=1}^{n}X_k e^{-r\tau_k}\Phi_k(h_i-\sigma\sqrt{\tau_i};\{\rho_{ij}\}) \\
&= e^{-r\tau_1}\int_{-\infty}^{h_1-\sigma\sqrt{\tau_1}}(S_{t_1}(z)-X_1)\phi(z)dz
\end{aligned}$$

where S_{t_1} is given by (D.15). Differentiating S_t with respect to V_t, we have

$$\begin{aligned}
\frac{\partial S_t}{\partial V_t} &= e^{-r\tau_1}\int_{-\infty}^{h_1-\sigma\sqrt{\tau_1}}\frac{\partial S_{t_1}(z)}{\partial V_t}\phi(z)dz + \frac{\partial h_1}{\partial V_t}\left[S_{t_1}(h_1-\sigma\sqrt{\tau_1})-X_1\right]\phi(h_1-\sigma\sqrt{\tau_1}) \\
&= e^{-r\tau_1}\int_{-\infty}^{h_1-\sigma\sqrt{\tau_1}}e^{(r-\frac{1}{2}\sigma^2)\tau_1+\sigma\sqrt{\tau_1}(-z)}\frac{\partial S_{t_1}(z)}{\partial V_{t_1}}\phi(z)dz
\end{aligned}$$

since $S_{t_1}(z) = X_1$ when $z = h_1 - \sigma\sqrt{\tau_1}$, and $V_{t_1} = V_t e^{(r-\frac{1}{2}\sigma^2)\tau_1+\sigma\sqrt{\tau_1}(-Z)}$. Assuming the theorem is true for $n-1$ payments, we assume $\frac{\partial S_{t_1}(z)}{\partial V_{t_1}} = \Phi_{n-1}(h'_{i+1}; \{\rho'_{ij}\})$ following the notation of the proof to Theorem D.4. Noting that $e^{-\frac{1}{2}\sigma^2\tau_1+\sigma\sqrt{\tau_1}(-z)}\phi(z) = \phi(z+\sigma\sqrt{\tau_1})$, and rewriting h'_{i+1}, we find

$$\frac{\partial S_t}{\partial V_t} = \int_{-\infty}^{h_1-\sigma\sqrt{\tau_1}} \phi(z+\sigma\sqrt{\tau_1})\Phi_{n-1}\left(\frac{h_{i+1} - \sqrt{\frac{\tau_1}{\tau_{i+1}}}(z+\sigma\sqrt{\tau_1})}{\sqrt{1-\frac{\tau_1}{\tau_{i+1}}}} ; \{\rho'_{ij}\}\right) dz.$$

Finally, we note this integral was solved in (D.17), giving

$$\frac{\partial S_t}{\partial V_t} = \Phi_n(h_i; \{\rho_{ij}\})$$

as required. If true for $n-1$ payments, the theorem is true for n payments, and since the theorem is true for one payment, by induction, the theorem is true for all n. □

Appendix E

Explaining the inverse leverage effect

The inverse leverage effect was seen in Table 4.2 in the situations where the firm has risk-free assets which completely offset the debt of the firm, and augment the exercise price of the call option. This situation was first described by Rubinstein (1983), with his option pricing formula a special case of (4.33). In this section, we give an intuitive explanation of how the inverse leverage effect (i.e. Black-Scholes overpricing in-the-money calls, and underpricing out-of-the-money calls) arises.

We consider the call prices on which the implied volatility ratios in Table 4.2 are based, and these are shown in Table E.1. In particular, we are interested in the behaviour of the call prices when risk-free assets are introduced to the firm's asset portfolio, without any change made to the instantaneous volatility of the risky assets. This effect is most simply examined when there is no debt, and therefore $S_t = V_t$ (an alternative would be when the firm has risk-free debt, and the effective exercise price of the option is increased). This corresponds to the upper left block of Tables 4.2 and E.1.

By ensuring all "firms" have $S_t = 10$ and $\sigma(S_t, t) = 0.4$, the true call price is a special case of (4.33), and is based on the Black-Scholes formula. Naïve pricing of the option (i.e. ignoring the possiblility of debt or heterogeneous assets) would use the Black-Scholes directly, with $S_t = 10$ and $\sigma = 0.4$, giving the option prices obtained when $\alpha_t = 1$ for the case of no debt. In general, these will not be equal to the true call prices, given by (4.33), unless of course the firm has homogeneous assets.

Because of the volatility matching process, the true call price is a complicated function of R_t and thus $\alpha_t \equiv \frac{V_t - R_t}{V_t} = \frac{S_t - R_t}{S_t}$, which features in three arguments of the

	No debt payment				Single debt payment				Two debt payments			
α_t	1	0.75	0.5	0.25	1	0.75	0.5	0.25	1	0.75	0.5	0.25
Strike	Leverage = 25%											
8	2.457	2.425	2.362	2.208	2.481	2.457	2.409	2.273	2.481	2.457	2.409	2.273
9	1.776	1.752	1.702	1.539	1.794	1.776	1.740	1.622	1.794	1.776	1.740	1.622
10	1.239	1.231	1.214	1.145	1.244	1.239	1.227	1.182	1.244	1.239	1.227	1.182
11	0.837	0.847	0.864	0.891	0.829	0.837	0.851	0.882	0.829	0.837	0.851	0.882
12	0.551	0.574	0.617	0.714	0.533	0.551	0.585	0.671	0.533	0.551	0.585	0.671
	Leverage = 50%											
8					2.503	2.489	2.457	2.362	2.505	2.489	2.457	2.362
9					1.810	1.799	1.776	1.702	1.811	1.799	1.776	1.702
10					1.249	1.245	1.239	1.214	1.248	1.245	1.239	1.214
11					0.822	0.827	0.837	0.864	0.820	0.826	0.837	0.864
12					0.517	0.527	0.551	0.617	0.514	0.527	0.551	0.617
	Leverage = 75%											
8					2.522	2.516	2.503	2.457	2.528	2.521	2.505	2.457
9					1.826	1.821	1.811	1.776	1.828	1.822	1.811	1.776
10					1.256	1.254	1.249	1.239	1.253	1.252	1.248	1.239
11					0.818	0.819	0.822	0.837	0.812	0.815	0.820	0.837
12					0.504	0.508	0.517	0.551	0.496	0.502	0.514	0.551

Table E.1. Call prices for options with time to maturity $\tau = 0.5$. All firms have $S_t = 10$ and $\sigma(S_t, t) = 0.40$. The single debt payment is at $\tau_2 = 2$, and the two debt payments are of identical size and made at $\tau_2 = 1$ and $\tau_3 = 1.5$. The leverage figure determines V_t, and this and $S_t = 10$ are used to find the required debt payment(s). The risk-free rate is $r = 0.05$ throughout.

Black-Scholes formula: the first (usually the price of the underlying asset), the second (usually the exercise price of the call) and the last (usually the constant volatility of the underlying asset). In the presence of risk-free assets, for firms aligned so that the stock has the same volatility regardless of $\alpha_t = \frac{S_t - R_t}{S_t}$, these three arguments are: $S_t - R_t = \alpha_t S_t$, $K - R_t e^{r\tau} = K - (1 - \alpha_t) S_t e^{r\tau}$ and $\sigma \frac{S_t}{S_t - R_t} = \frac{\sigma}{\alpha_t}$ respectively, and the call price is given by

$$C_t = \alpha_t S_t \Phi(\omega_t) - \left[K e^{-r\tau} - (1 - \alpha_t) S_t\right] \Phi(\omega_t - \tfrac{\sigma}{\alpha_t}\sqrt{\tau}) \tag{E.1}$$

where we are assuming the special case of no debt (giving $V_t = S_t$), with

$$\omega_t = \frac{\ln \alpha_t S_t - \ln(K - (1 - \alpha_t) S_t e^{r\tau}) + (r + \frac{1}{2}\frac{\sigma^2}{\alpha_t^2})\tau}{\frac{\sigma}{\alpha_t}\sqrt{\tau}}.$$

When we increase the proportion of firm value held in risky assets, α_t (through a decrease in R_t), the first argument increases, which *ceteris paribus* increases the call price; the second argument also increases, however this decreases the call price; and the last argument decreases, also decreasing call price. What we see in practice depends on the size of K.

As seen in Table E.1, when K is small, so that the call is in-the-money, call price increases with α_t and Black-Scholes overvalues the call. However, when K is large,

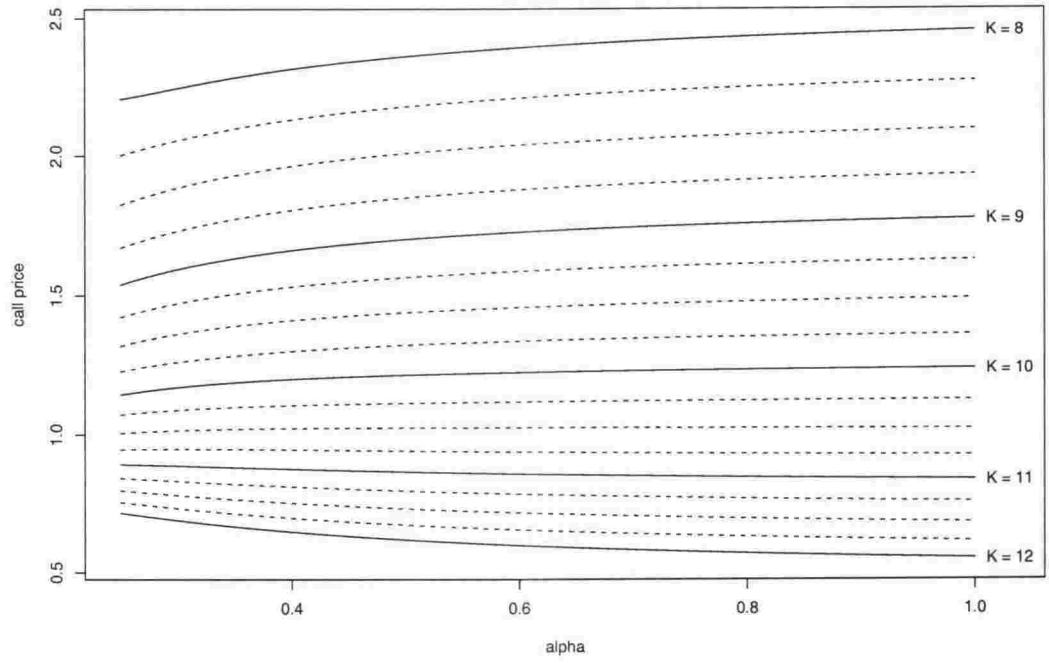

Figure E.1. Displaced diffusion call prices for options with time to maturity $\tau = 0.5$. All firms have $S_t = 10$ and $\sigma(S_t, t) = 0.40$, but a variable α_t. Each contour is for a fixed exercise price, and is a plot of the call price C_t against α_t, and are plotted for K at intervals of 0.25, with the integer values indicated. The risk-free rate is $r = 0.05$.

so that the call is out-of-the-money, as α_t increases, call price decreases, and the Black-Scholes price is too low. Calculation of

$$\frac{\partial C_t}{\partial \alpha_t} = S_t \left[\Phi(\omega_t) - \Phi(\omega_t - \tfrac{\sigma}{\alpha_t}\sqrt{\tau}) \right] - \frac{\sigma\sqrt{\tau}}{\alpha_t^2} \left[Ke^{-r\tau} - (1-\alpha_t)S_t \right] \phi(\omega_t - \tfrac{\sigma}{\alpha_t}\sqrt{\tau})$$

reveals that the sign of this derivative depends on a complicated function of the parameters. Nonetheless, it can easily be plotted for any choice of parameters. Generally the call price function (E.1) is monotonic, as shown in Figure E.1; however it is not necessarily so. The call price contour for exercise price $K = 10.5$ is one such case, and this is shown in Figure E.2. As seen, the range of call prices is very small; however the non-monotonicity is clear.

Regardless of the relationship with α_t, the relationship between implied volatility and strike price for any fixed $\alpha_t < 1$ is consistent: an increasing relationship, and this is shown in Figure E.3. Since this relationship is non-standard (compared to the traditional leverage effect discussed in Section 4.1.4) we refer to it as the inverse leverage effect.

How the inverse leverage effect arises becomes clear when we examine the probability density functions for share price at exercise, $f_{S_{t+\tau}}(s)$, for the aligned processes (with

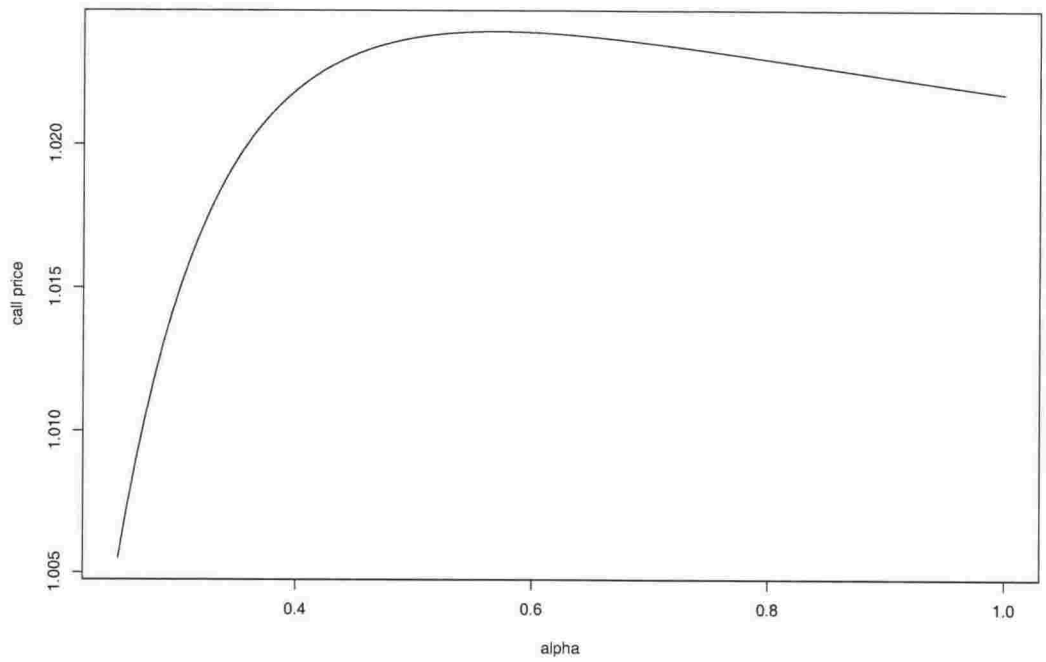

Figure E.2. Displaced diffusion call prices for options with exercise price $K = 10.5$ and time to maturity $\tau = 0.5$. All firms have $S_t = 10$ and $\sigma(S_t, t) = 0.40$, but a variable α_t and we plot the call price C_t against α_t. The risk-free rate is $r = 0.05$.

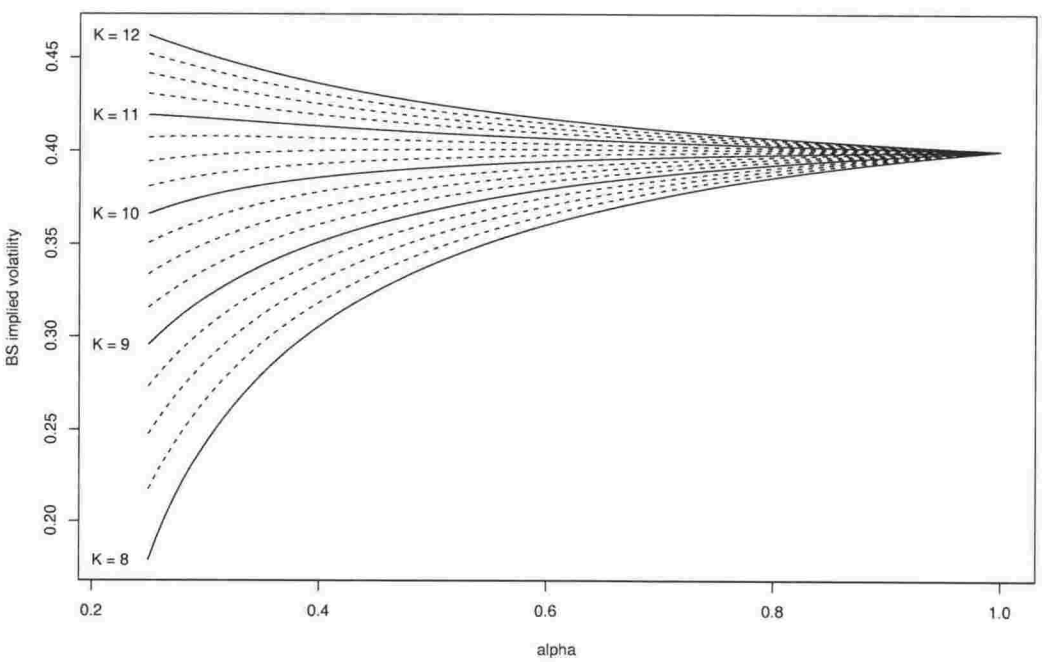

Figure E.3. Displaced diffusion Black-Scholes implied volatilities for options with time to maturity $\tau = 0.5$. All firms have $S_t = 10$ and $\sigma(S_t, t) = 0.40$, but a variable α_t. Each contour is for a fixed exercise price, and is a plot of the BS implied volatility against α_t. These are plotted for K at intervals of 0.25, with the integer values indicated. The risk-free rate is $r = 0.05$.

and without non-risky assets), and also the function $\max(s-K,0)f_{S_{t+\tau}}(s)$ which is integrated to yield the call price.

Theorem E.1 (Displaced diffusion density function) *The density function for a displaced diffusion process, defined by $S_{t+\tau} = V_{t+\tau}$ and with V_t specified in (4.27) and (4.26), given S_t, is given by*

$$f_{S_{t+\tau}}(s) = \begin{cases} f_{A_{t+\tau}}\big(s-(1-\alpha_t)S_t e^{r\tau}\big) & s > (1-\alpha_t)S_t e^{r\tau} \\ 0 & s \leq (1-\alpha_t)S_t e^{r\tau} \end{cases}$$

where $A_{t+\tau}$ is a lognormal random variable with parameters $\ln \alpha_t S_t + (r - \frac{1}{2}\frac{\sigma^2}{\alpha_t^2})\tau$ and $\sigma^2\tau/\alpha_t^2$.

Proof From (4.26) we note that

$$S_{t+\tau} = A_{t+\tau} + (1-\alpha_t)S_t e^{r\tau}$$

and since A_t is a GBM process with volatility parameter σ/α_t, given A_t, $A_{t+\tau}$ is a lognormal random variable with parameters $\ln \alpha_t S_t + (r - \frac{1}{2}\frac{\sigma^2}{\alpha_t^2})\tau$ and $\sigma^2\tau/\alpha_t^2$. Thus, for $s > (1-\alpha_t)S_t e^{r\tau}$

$$F_{S_{t+\tau}}(s) = P(S_{t+\tau} < s) = P\big(A_{t+\tau} < s - (1-\alpha_t)S_t e^{r\tau}\big) = F_{A_{t+\tau}}\big(s-(1-\alpha_t)S_t e^{r\tau}\big)$$

where $F_{A_{t+\tau}}(x)$ is the cdf of $A_{t+\tau}$ given A_t, and for $s < (1-\alpha_t)S_t e^{r\tau}$, $F_{S_{t+\tau}}(s) = 0$. Differentiating the cdf with respect to s, we obtain

$$f_{S_{t+\tau}}(s) = \begin{cases} f_{A_{t+\tau}}\big(s-(1-\alpha_t)S_t e^{r\tau}\big) & s > (1-\alpha_t)S_t e^{r\tau} \\ 0 & s \leq (1-\alpha_t)S_t e^{r\tau} \end{cases}$$

where $f_{A_{t+\tau}}(x)$ is the pdf of $A_{t+\tau}$ given A_t, as required. □

The density function for a geometric Brownian motion process $S_{t+\tau}$, given $S_t = 10$, $\sigma = 0.4$ and $\tau = 0.5$, is shown by the solid line, in Figure E.4. Under the GBM assumption, corresponding to $\alpha_t = 1$ in the case of no debt, the share price at exercise $S_{t+\tau}$ is a lognormal random variable, with parameters $\ln S_t + (r - \frac{1}{2}\sigma^2)\tau$ and $\sigma^2\tau$. Also shown are the density functions for two displaced diffusion processes, both with $S_t = 10$ and $\sigma(S_t, t) = 0.4$, and with $\alpha_t = 0.75$ and $\alpha_t = 0.5$ (shown by the dashed and dotted lines respectively). The form of the density function for $S_{t+\tau}$ in the displaced diffusion case is related to the lognormal, and is given in Theorem E.1. While the GBM process has a small probability of falling close to zero,

the displaced diffusion processes are always at least the size of the risk-free assets $(1-\alpha_t)S_t e^{r\tau}$. In addition, the mode of the density moves to the left as α_t increases, and the right tail becomes heavier. In particular, the GBM density overestimates the probability of a sharp decrease in stock price, underestimates the probability of a small decrease or small increase, overestimates the probability of a moderate increase, and underestimates the probability of a large increase.

In order to see how the inverse leverage effect results, it is useful to examine the plot of $\max(0, s-K)f_{S_{t+\tau}}(s)$. The call option prices are obtained by integrating this function over positive values of s and discounting, i.e.

$$C_t = e^{-r\tau}\int_{s=0}^{\infty} \max(0, s-K)f_{S_{t+\tau}}(s)ds. \tag{E.2}$$

The integrand of (E.2) is plotted in Figure E.5 for $\alpha_t \in \{1, 0.75, 0.5\}$ and for $K \in \{8, 12\}$. In the left-hand plot of Figure E.4, corresponding to the exercise price of $K = 8$, for small values of $s > K$, the function is greatest in the GBM case; however at the tail, the function is greater when $\alpha_t < 1$. The overall effect is demonstrated through the integrals evaluated in Table E.1, where we see the call prices are 2.457, 2.425 and 2.362 for $\alpha_t = 1$, 0.75 and 0.5 respectively. Returning to Figure E.4, we conclude the additional area for the GBM case at the maximum of the integrand outweighs the area lost at the tail. Thus, when the call is in-the-money and the stock price is a displaced diffusion process with $\alpha_t < 1$, the Black-Scholes formula overprices the call option, due to the fact that it overestimates the probability of a moderate positive movement in stock price.

In the right-hand plot of Figure E.4, corresponding to the exercise price of $K = 12$, we see a different effect. In this case, again the heavier tail of the density functions for $\alpha_t < 1$ is amplified, however the effect at the mode of the probability densities is not as great, since K is too large. Once again, the overall effect is found in Table E.1, where we see the call prices are 0.551, 0.574, and 0.617 for $\alpha_t = 1$, 0.75 and 0.5 respectively. Returning to Figure E.4, we conclude the additional area for the GBM case for small s is outweighed by the area lost at the tail. Thus, when the call is out-of-the-money and the stock price follows a displaced diffusion process with $\alpha_t < 1$, the Black-Scholes formula underprices the call option, due to the fact that it underestimates the probability of a gross upward movement in stock price, an effect which in this case is more pronounced than the overestimated probability of a moderate increase. This probability has arisen, because the risky assets have an inflated volatility in order to achieve $\sigma(S_t, t) = 0.4$, i.e., they have volatility σ/α_t.

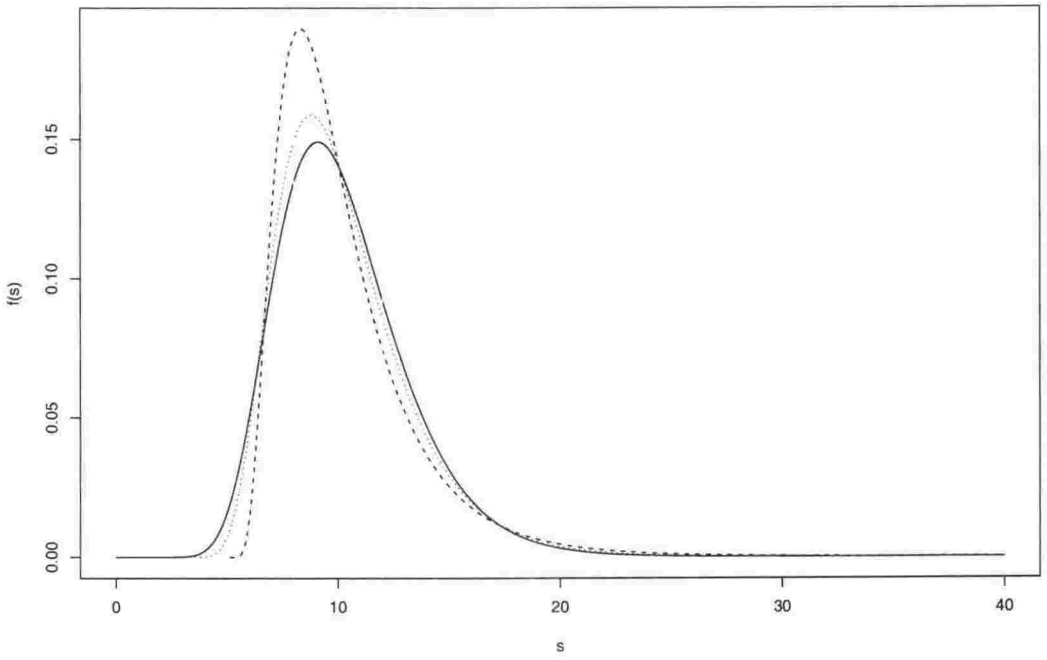

Figure E.4. Displaced diffusion density functions for $S_{t+\tau}$, with $S_t = 10$, $\sigma = 0.4$, $\tau = 0.5$ and $r = 0.05$. The solid line is for the GBM process with $\alpha_t = 1$; the dashed line for the process with $\alpha_t = 0.75$ and the dotted line for the process with $\alpha_t = 0.5$. The vertical lines plotted in grey, correspond to exercise prices $K = 8$ and 12.

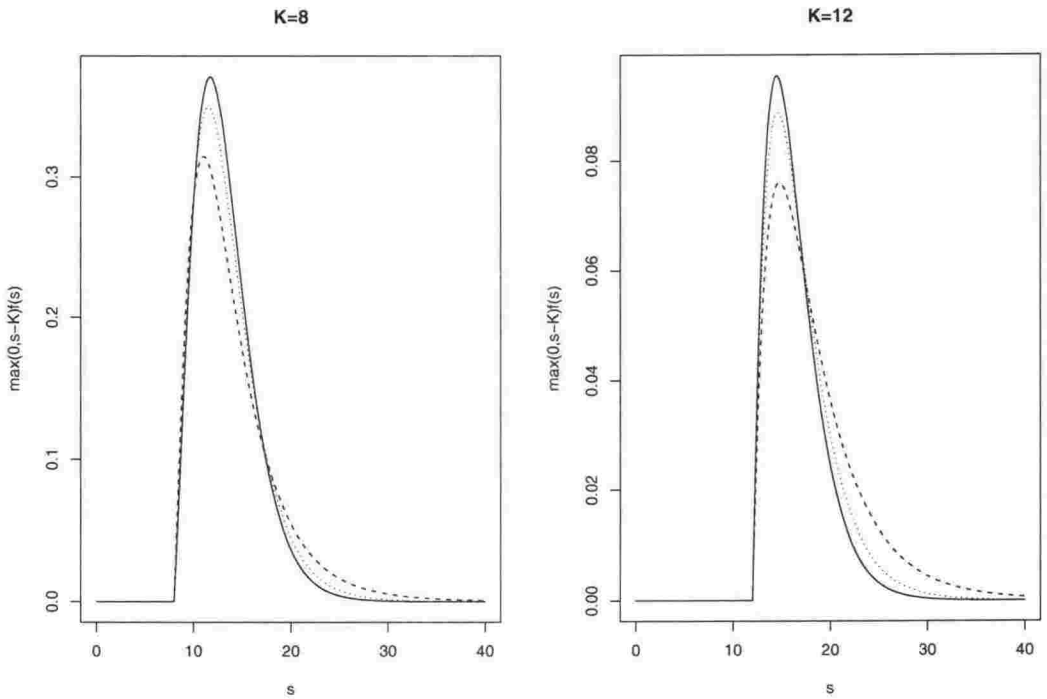

Figure E.5. Displaced diffusion call prices are obtained by integrating the functions shown: $\max(0, s - K) f_{S_{t+\tau}}(s)$ from $s = 0$ to ∞, corresponding to the density functions $f_{S_{t+\tau}}(s)$ shown in Figure E.4. The solid line is for the GBM process with $\alpha_t = 1$; the dashed line for the process with $\alpha_t = 0.75$ and the dotted line for the process with $\alpha_t = 0.5$. Stock price is $S_t = 10$, $\sigma = 0.4$, $\tau = 0.5$ and $r = 0.05$. The plot on the left has $K = 8$ and the call is in-the-money. The plot on the right has $K = 12$ and the call is out-of-the-money.

These effects flow through to the Black-Scholes implied volatilities on which the ratios seen in Table 4.2 are based, resulting in a positive relationship between exercise price and implied volatility, which is the inverse of the traditional leverage effect.

Bibliography

Abramowitz, M. & Stegun, I. A., eds (1968), *Handbook of Mathematical Functions*, Dover Publications, Inc.

Aït-Sahalia, Y. (1996), 'Nonparametric pricing of interest rate derivative securities', *Econometrica* **64**(3), 527–560.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2001), 'The distribution of realized exchange rate volatility', *Journal of the American Statistical Association* **96**, 42–55.

Anderson, M. & Grier, D. A. (1992), 'Robust, non-parametric measures of exchange rate variability', *Applied Economics* **24**, 951–958.

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W. & Tukey, J. (1972), *Robust estimates of location: survey and advances*, Princeton University Press.

Barndorff-Nielsen, O. E. & Shephard, N. (2001), 'Non-Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial economics', *Journal of the Royal Statistical Society, Series B* **63**, 167–241.

Bates, D. S. (2000), 'Post-'87 crash fears in the S&P 500 futures option market', *Journal of Econometrics* **94**, 181–238.

Beckers, S. (1983), 'Variances of security price returns based on high, low and closing prices.', *Journal of Business* **56**(1), 97–112.

Black, F. (1976), 'Studies of stock price volatility changes', *Proceedings of the 1976 meetings of the American Statistical Association, business and economic statistics section* pp. 177–181.

Black, F. & Scholes, M. (1973), 'The pricing of options an corporate liabilities', *Journal of Political Economy* **81**(3), 637–654.

Blattberg, R. & Gonedes, N. (1974), 'A comparison of the stable and Student distributions as statistical models for stock prices', *Journal of Business* **47**, 244–280.

Bollerslev, T., Chou, R. Y. & Kroner, K. F. (1992), 'ARCH modeling in finance', *Journal of Econometrics* **52**, 5–59.

Bollerslev, T. & Mikkelson, H. O. (1996), 'Modeling and pricing long memory in stock market volatility', *Journal of Econometrics* **73**, 151–184.

Chen, G. & Ryan, P. J. (1996), 'Displaced diffusion option pricing with two risky assets', *Journal of Interdisciplinary Economics* **7**, 205–216.

Christie, A. A. (1982), 'The stochastic behaviour of common stock variances', *Journal of Financial Economics* **10**, 407–432.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Cleveland, W. S., Grosse, E. & Shyu, W. M. (1992), *Statistical Models in S*, Wadsworth and Brooks/Cole, chapter 8: Local regression models, pp. 309–376.

Cohen, M. (1991), *Configural Polysampling*, John Wiley and Sons, Inc, chapter 2: The background of configural polysampling: a historical perspective.

Cox, J. C. (1996), 'The constant elasticity of variance option pricing model', *The Journal of Portfolio Management* **Special Issue**, 15–17.

Cox, J. C. & Ross, S. A. (1976), 'The valuation of options for alternative stochastic processes', *Journal of Financial Economics* **3**, 145–166.

Curnow, R. & Dunnett, C. (1962), 'The numerical evaluation of certain multivariate normal integrals', *The Annals of Mathematical Statistics* **33**, 571–579.

Datastream (2002), 'Datastream Advance 3.5', Thomson Financial Limited. www.thomsonfinancial.com.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Ding, Z. & Granger, C. W. (1996), 'Modelling volatility persistence of speculative returns: A new approach', *Journal of Econometrics* **73**, 185–215.

Ding, Z., Granger, C. W. & Engle, R. (1993), 'A long memory property of stock market returns and a new model', *Journal of Empirical Finance* **1**, 83–106.

Emanuel, D. C. & MacBeth, J. D. (1982), 'Further results on the constant elasticity of variance call option pricing model', *Journal of Financial and Quantitative Analysis* **17**(4), 533–554.

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

Fama, E. (1965), 'The behaviour of stock market prices', *Journal of Business* **38**, 34–105.

Figlewski, S. (1997), 'Forecasting volatility', *Financial Markets, Institutions & Instruments* **6**(1), 1–88.

Frey, R. & Sommer, D. (1998), 'The generalization of the Geske-formula for compound options to stochastic interest rates is not trivial - a note', *Journal of Applied Probability* **35**(2), 501–509.

Garman, M. B. & Klass, M. J. (1980), 'On the estimation of security price volatilities from historical data', *Journal of Business* **53**(1), 67–78.

Genz, A. (1992), 'Numerical computation of multivariate normal probabilities', *Journal of Computational and Graphical Statistics* **1**, 141–150.

Genz, A. (1993), 'Comparison of methods for the computation of multivariate normal probabilities', *Computing Science and Statistics* **25**, 400–405.

Geske, R. (1977), 'The valuation of corporate liabilities as compound options', *Journal of Financial and Quantitative Analysis* **12**, 541–552.

Geske, R. (1979), 'The valuation of compound options', *Journal of Financial Economics* **7**, 63–81.

Goodall, C. (2000), *Understanding Robust and Exploratory Data Analysis*, Wiley Classics Library Edition, chapter 11: M-estimators of location: an outline of the theory, pp. 339–403.

Gourieroux, C. & Jasiak, J. (2001), *Financial Econometrics: Problems, Models and Methods*, Princeton University Press.

Granger, C. W. J. & Ding, Z. (1995), 'Some properties of absolute returns, an alternative measure of risk', *Annales d'économie et de statistique* **40**, 67–91.

Gray, A. G. & Thomson, P. J. (1990), 'Invited commentary on 'STL: A seasonal-trend decomposition procedure based on loess' by R.B. Cleveland et al.', *Journal of Official Statistics* **6**(1), 47–54.

Graybill, F. A. (1976), *Theory and Application of the Linear Model*, Duxbury Press.

Gross, A. M. (1976), 'Confidence interval robustness with long-tailed symmetric distributions', *Journal of the American Statistical Association* **71**, 409–416.

Hamilton & Susmel (1994), 'Autoregressive conditional heteroskedasticity and changes in regime', *Journal of Econometrics* **64**, 307–333.

Harrison, J. M. & Pliska, S. R. (1981), 'Martingales and stochastic integrals in the theory of continuous trading', *Stochastic Processes and their Applications* **11**, 215–260.

Harvey, A., Ruiz, E. & Shephard, N. (1994), 'Multivariate stochastic variance models', *Review of Economic Studies* **61**, 247–264.

Hoaglin, D., Mosteller, F. & Tukey, J., eds (2000), *Understanding Robust and Exploratory Data Analysis*, Wiley Classics Library Edition.

Huber, P. J. (1981), *Robust Statistics*, John Wiley and Sons.

Hull, J. C. (1997), *Options, Futures, and Other Derivatives*, 3rd edn, Prentice-Hall, Inc.

Hull, J. C. (2000), *Options, Futures, and Other Derivatives*, 4th edn, Prentice-Hall, Inc.

Hurst, S. & Platen, E. (1997), *L_1-Statistical Procedures and Related Topics*, IMS Lecture Notes–Monograph Series 31, California, chapter : The marginal distribution of returns and volatility.

Iglewicz, B. (2000), *Understanding Robust and Exploratory Data Analysis*, Wiley Classics Library Edition, chapter 12: Robust scale estimators and confidence intervals for location, pp. 404–431.

Iglewicz, B. & Martinez, J. (1982), 'Outlier detection using robust measures of scale', *Journal of Statistical Computation and Simulation* **15**, 285–293.

Ihaka, R. & Gentleman, R. (1996), 'R: A language for data analysis and graphics', *Journal of Computational and Graphical Statistics* **5**(3), 299–314.

JPMorgan (1996), 'RiskMetrics', Technical Document, 4th ed.

Kafadar, K. (1982), 'A biweight approach to the one-sample problem', *Journal of the American Statistical Association* **77**, 416–424.

Kunitomo, N. (1992), 'Improving the Parkinson method of estimating security price volatilities', *Journal of Business* **65**(2), 295–302.

Lamoureux, C. & Lastrapes, W. (1990), 'Persistence in variance, structural change, and the GARCH model', *Journal of Business and Economic Statistics* **8**, 225–234.

Lax, D. A. (1985), 'Robust estimators of scale: finite-sample performance in long-tailed symmetric distributions', *Journal of the American Statistical Association* **80**, 736–741.

Lehmann, E. & Casella, G. (1998), *Theory of Point Estimation*, Springer-Verlag New York Ltd.

Leland, H. E. (1994), 'Corporate debt value, bond covenants, and optimal capital structure', *Journal of Finance* **49**, 1213–1252.

Liesenfeld, R. & Jung, R. C. (2000), 'Stochastic volatility models: conditional normality versus heavy-tailed distributions', *Journal of Applied Econometrics* **15**, 137–160.

MacBeth, J. D. & Merville, L. J. (1980), 'Tests of the Black-Scholes and Cox call option valuation models', *Journal of Finance* **35**(2), 285–300.

Martinez, J. & Iglewicz, B. (1981), 'A test for departure from normality based on a biweight estimator of scale', *Biometrika* **68**, 331–333.

Mayhew, S. (1995), 'Implied volatility', *Financial Analysts Journal* **51**(4), 8–20.

McConnell, M. & Perez-Quiros, G. (2000), 'Output fluctuations in the United States: What has changed since the early 1980's?', *The American Economic Review* **90**, 1464–1476.

McLachlan, G. J. & Krishnan, T. (1997), *The EM algorithm and extensions*, Wiley Series in Probability and Statistics.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974), *Introduction to the theory of statistics*, 3rd edn, McGraw-Hill.

Morgenthaler, S. & Tukey, J. W., eds (1991), *Configural Polysampling: A Route to Practical Robustness*, John Wiley & Sons, Inc.

Officer, R. R. (1973), 'The variability of the market factor of the New York Stock Exchange', *Journal of Business* **46**(3), 434–453.

Parkinson, M. (1980), 'The extreme value method for estimating the variance of the rate of return.', *Journal of Business* **53**(1), 61–65.

Randal, J. A. (1998), The constant elasticity of variance option pricing model, MSc Thesis, Victoria University of Wellington.

Rogers, W. H. & Tukey, J. W. (1972), 'Understanding some long-tailed symmetrical distributions', *Statistica Neerlandica* **26**, 211–226.

Rosenberger, J. L. & Gasko, M. (2000), *Understanding Robust and Exploratory Data Analysis*, Wiley Classics Library Edition, chapter 10: Comparing location estimators: trimmed means, medians, and trimean, pp. 297–338.

Rousseeuw, P. J. & Croux, C. (1993), 'Alternatives to the median absolute deviation', *Journal of the American Statistical Association* **88**, 1273–1283.

Rubinstein, M. (1983), 'Displaced diffusion option pricing', *Journal of Finance* **38**, 213–217.

Rydén, T., Teräsvirta, T. & Åsbrink, S. (1998), 'Stylized facts of daily return series and the hidden Markov model', *Journal of Applied Econometrics* **13**, 217–244.

Schroder, M. (1989), 'Computing the constant elasticity of variance option pricing formula', *Journal of Finance* **44**(1), 211–219.

Seber, G. (1977), *Linear Regression Analysis*, Wiley Series in Probability and Mathematical Statistics.

Shalit, H. & Yitzhaki, S. (1984), 'Mean-Gini, portfolio theory, and the pricing of risky assets', *Journal of Finance* **39**, 1449–1468.

Shephard, N. (1996), *Time Series Models in Econometrics, Finance and Other Fields*, Chapman and Hall, London, chapter : Statistical aspects of ARCH and stochastic volatility.

Simon, G. (1976), 'Computer simulation swindles, with applications to estimates of location and dispersion', *Applied Statistics* **25**(3), 266–274.

Taylor, S. (1986), *Modelling Financial Time Series*, John Wiley & Sons.

Toft, K. B. & Prucyk, B. (1997), 'Options on leveraged equity: theory and empirical tests', *Journal of Finance* **52**, 1151–1180.

Turner, A. L. & Weigel, E. J. (1992), 'Daily stock market volatility: 1928-1989', *Management Science* **38**(11), 1586–1609.

Venables, W. N. & Ripley, B. (1999), *Modern Applied Statistics with S-PLUS*, 3rd edn, Springer.

Yatrakos, Y. G. (1991), 'A note on Tukey's polyefficiency', *Biometrika* **78**(3), 702–703.