Abnormal Returns, Risk, and Options in Large Data Sets

Silvia Caserta
Tinbergen Institute Rotterdam

Jón Daníelsson
London School of Economics and IoES at University of Iceland

Casper G. de Vries*
Tinbergen Institute Rotterdam and Erasmus University Rotterdam

Abstract

Large data sets in finance with millions of observations have become widely available. Such data sets enable the construction of reliable semi-parametric estimates of the risk associated with extreme price movements. Our approach is based on semi-parametric statistical extreme value analysis, and compares favorably with the conventional finance normal distribution based approach. It is shown that the efficiency of the estimator of the extreme returns may benefit from high frequency data. Empirical tail shapes are calculated for the German Mark - US Dollar foreign exchange rate, and we use the semi-parametric tail estimates in combination with the empirical distribution function to evaluate the returns on exotic options.

Keywords: Extreme value theory, tail estimation, high frequency data, exotic options.

*Corresponding author: C.G. de Vries, Tinbergen Institute Rotterdam, P.O. Box 1738, 3000 Rotterdam, The Netherlands, e-mail *cdevries@few.eur.nl*. Related papers can be downloaded from *http://www.hag.hi.is/∼ jond/research*. We benefited from the discussion at the conference on "The Statistical Analysis of Large Data Sets in Business and Economics" organized by RIBES. The comments and suggestions by two anonymous referees are gratefully acknowledged. We like to thank Terry Cheuk for his valuable advise concerning the barrier option, Winfred Hallerbach for his perceptive comments and the Olsen & Associates company for making their data available. Daníelsson was supported by the research contribution of the Icelandic banks.

1 Introduction

The size of available data sets in economics and finance has grown rapidly. Initially, most empirically oriented economists worked with data sets smaller than a few hundred observations. During the seventies the size of available data sets in finance and marketing grew into the thousands, and nowadays millions of observations per series are commonplace. For example the Olsen & Associates company has over ten millions observations of screen quotes for the DM-USD contract. In this paper we discuss some of the exiting statistical and economic issues that can be dealt with given such voluminous data sets.

We are concerned with the return, i.e. percentage gain or loss, on either a single financial asset or on a portfolio (linear combinations) of different assets. In finance it is often assumed for convenience that returns are normally distributed. For certain questions this assumption is harmless and expedient. But it is also well-known that the empirical distribution of asset returns exhibits many more tail realizations than could be expected on the basis of the normal model. This so called heavy tail feature of the return data has become a stylized fact among applied researchers, see Ballie and McMahon (1989, p. 135) and Campbell, Lo and MacKinlay (1997, ch. 1). Since the normal model leads to an underprediction of tail events, for applications in areas like risk management (which focuses on the probability of bankruptcy), it is more prudent to work with the heavy tailed distributional assumption.

It is a fortunate coincidence that all heavy tailed distributions exhibit, to a first order approximation, the same hyperbolic tail behaviour. Due to this property it suffices to use a semi-parametric analysis for questions concerning tail risk, where only the tails of the return distribution are modeled parametrically. In the theory section it is argued that the mean-squared error (MSE) property of the semi-parametric estimator of the hyperbolic tail coefficient benefits from an increase in the frequency of the data . We also argue that the estimation procedure requires a large number of observations to begin with, because it relies on bootstrap resamples which have to be of smaller order than the original sample size, but still contain a sufficient number of extreme realizations.

In the empirical section we first demonstrate the benefit of an increase in the frequency of the returns on the hyperbolic tail shape parameter estimates. Secondly, we show how a sizeable data set can be exploited to compute the returns on derivatives conditional on the underlying returns being heavy tailed by a semi-parametric method. Conventional analysis uses a small number of observations and proceeds on the basis of normality. Here the size of the data set allows us to use the empirical distribution in the middle and the parametric tail estimates at the two ends of the distribution itself. The effects of non-normality are shown to be important for pricing of an exotic option. While the first part of the paper is a review of existing results, the analysis of the exotic option is new.

2 Theoretical Benefit of High Frequency Data

It is a stylized fact in finance that asset returns are fat tailed distributed. This can be modeled by assuming that the return distribution has regularly varying tails. Formally, a distribution function $F(x)$ varies regularly at infinity if

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-\alpha}, \quad \alpha > 0, \quad x > 0 \tag{1}$$

where α is called the tail index. The tail index indicates the thickness of the tails and equals the number of bounded moments.

From (1) it follows that to a first order approximation at infinity all heavy tailed distributions have the same Pareto tail shape $a\,x^{-\alpha}$, where a is a scaling constant, see Feller (1971, VIII.8). Hence, if one is only interested in the tail behavior such as in the question of the bankruptcy probability, one does not need to model a specific distribution $F(x)$. Instead, one can proceed semiparametrically by estimating $a\,x^{-\alpha}$ only.

The tail index can be estimated by the Hill (1975) estimator

$$\widehat{\frac{1}{\alpha}} = \frac{1}{m}\sum_{i=1}^{m} log \frac{X_{(i)}}{X_{(m+1)}} \tag{2}$$

where the $X_{(i)}$ are the largest descending order statistics $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(m)} \geq X_{(m+1)} \geq \cdots \geq X_{(n)}$, pertaining to the sample $X_1, \ldots, X_n$. The Hill estimator is motivated by the maximum likelihood estimator for the hyperbolic coefficient of the Pareto distribution: Above an appropriately chosen high threshold $X_{(m+1)}$, the tail of a regularly varying distribution function $F(x)$ is approximately in conformity with the shape of the Pareto distribution, see also De Haan (1990).

Once the tail index has been estimated, large quantile estimates can be obtained as follows. Consider two tail probabilities p and t. For example take $p < 1/n < t \leq m/n$. Let x_p and x_t denote the corresponding quantiles. Hence, $p \approx a\,x_p^{-\alpha}$ and $t \approx a\,x_t^{-\alpha}$. Combining these two expressions yields $x_p \approx x_t\,(t/p)^{1/\alpha}$. Replace x_t by $X_{(m+1)}$ and use the empirical distribution function for $t \approx m/n$. The quantile estimator reads

$$\hat{x}_p = X_{(m+1)} \left(\frac{m}{np}\right)^{\widehat{1/\alpha}} \tag{3}$$

see, e.g. De Haan, Jansen, Koedijk and de Vries (1994). Note that the statistical properties of $\hat{x}_p$ are in essence determined by the properties of $\widehat{1/\alpha}$.

The theory for the statistics $\widehat{1/\alpha}$ and $\hat{x}_p$ is derived under the assumption of independence. Typically, financial returns exhibit the fair game property. The reason is a simple arbitrage argument. If one knew that the asset price was going to be high tomorrow, one would buy today. But this already raises today's price and quickly eradicates profit opportunities, leaving the expected net return equal to zero (net of the growth rate of the firm or the economy). Hence, returns data typically do not exhibit dependence in the first moment. The second moments, however, do display dependence. Markets happen to have periods of quiescence and turbulence. A popular model for this feature is the class of ARCH processes, see Engle (1982). The ARCH model relates current variance to the past squared realizations. Kearns and Pagan (1997) considered how this second moment dependence might influence the properties of the statistics (2) and (3). De Haan et al. (1989) showed that the stationary distribution of the ARCH process varies regularly at infinity; and Resnick and Starica (1998) proved the consistency of the estimators under the assumption of ARCH.

Both estimator (2) and (3) depend on a judiciously chosen threshold index m. To tackle this issue, consider the following second order tail expansion

$$F(x) = 1 - ax^{-\alpha}\left[1 + b\,x^{-\beta} + o(x^{-\beta})\right], \qquad \text{as} \quad x \to \infty \tag{4}$$

with $\beta > 0, a > 0$ and $b \in \Re$. This expansion applies to a large subclass of the heavy tailed distributions, see De Haan and Stadtmüller (1996). For example the Student-t, the symmetric sum-stable distributions with characteristic exponent $\alpha \in (1,2)$ and type II extreme value distributions, all admit (4). For this class of distributions the bias and the variance of $\widehat{1/\alpha}$ are straightforward to calculate. The asymptotic mean squared error (AMSE) is then found as

$$\text{AMSE}\left(\widehat{\frac{1}{\alpha}}\right) = \text{E}\left[\left(\widehat{\frac{1}{\alpha}} - \frac{1}{\alpha}\right)^2\right] = \frac{1}{\alpha^2}\frac{\beta^2\,b^2}{(\alpha+\beta)^2}a^{-\frac{2\beta}{\alpha}}\left(\frac{m}{n}\right)^{\frac{2\beta}{\alpha}} + \frac{1}{\alpha^2}\frac{1}{m} \tag{5}$$

In the AMSE sense, it is optimal to have the bias and the variance vanish at the same rate as $n \to \infty$, otherwise one of the two will dominate the other, and this brings down the rate of convergence of the AMSE below the best obtainable rate. Hall (1982) first showed that this optimal rate determines the optimal number of order statistics as follows

$$\overline{m} = c\,n^{\frac{2\,\beta}{(2\,\beta+\,\alpha)}}, \qquad c > 0 \tag{6}$$

where c depends on the parameters a, b, α, β (this follows from equating the partial derivative of the AMSE$(\widehat{1/\alpha})$ with respect to m to 0). Hall(1982) and Goldie and Smith (1987) showed that when $m = \overline{m}(n)$, the rescaled Hill estimator $\sqrt{\overline{m}}\widehat{1/\alpha}$ is asymptotically normally distributed. A similar result applies to the quantile estimator $\hat{x}_p$.

We will now give an example of how an increase in the frequency of the data improves the efficiency of the tail estimators. The example is based on Dacorogna, Müller, Pictet and de Vries (1995). Suppose that the data are generated by a symmetric sum-stable distribution with characteristic exponent $\alpha \in (1,2)$; note that the characteristic exponent equals the tail index when $\alpha \neq 2$. For the stable distributions $\beta = \alpha$ in (4), the scaling constants can be easily obtained from Feller (1971, ch. XVII.7). In this case the expression for the AMSE in (5) at $m = \bar{m}$ becomes

$$\mathrm{AMSE}_n = \frac{3}{\alpha^2}\left(\frac{b}{4a}\right)^{\frac{2}{3}} n^{-\frac{2}{3}} \tag{7}$$

Instead of considering single period returns, now consider w-period returns. A w-period return is simply the sum of w consecutive individual returns. For the w-convoluted data, by using the additivity property of stable variables, the AMSE reads

$$\mathrm{AMSE}_{\frac{n}{w}} = \frac{3}{\alpha^2}\left(\frac{b}{4a}\right)^{\frac{2}{3}}\left(\frac{n}{w}\right)^{-\frac{2}{3}} \tag{8}$$

It follows that the time aggregated data yield less efficient estimates of $1/\alpha$.

Another benefit from large data sets is that these allow one to locate the optimal $\overline{m}$ in a statistical satisfactory way. While the theoretical expression for $\overline{m}$ was given in (6), we did not explain how this number can be obtained in practice such that the asymptotic normality is preserved. This problem was recently solved by Daníelsson, De Haan, Peng and de Vries (1997) and Daníelsson and de Vries (1998) by means of a bootstrap procedure. The idea is to construct the bootstrap expectation of $(\widehat{1/\alpha} - 1/\alpha)^2$, and to minimize it with respect to m.

Two hurdles have to be crossed before this procedure can be used. The first problem is that the benchmark $1/\alpha$ is unknown. A solution to this problem is to replace $1/\alpha$ by an alternative estimator, with the same rate of convergence as $\widehat{1/\alpha}$ but different multiplicative constant, in the minimization of the $\mathrm{AMSE}(\widehat{1/\alpha})$. The second problem is that the bootstrap procedure applied to the entire sample generates a threshold m which only converges in distribution to the optimal level. To achieve a convergence in probability, a subsample bootstrap technique must be employed. To be able to implement such a procedure one needs to construct subsamples which are, on the one hand, an order of magnitude smaller than the full sample size. On the other hand, because the outliers are rare by their very nature, one needs subsamples which are still quite sizable. Hence the use of large data sets, as the Olsen & Associates data set, allows one to exploit the subsample bootstrap procedure. As Daníelsson et al. (1997) and Daníelsson and de Vries (1998) suggest the sample size n needs at least to be in the order of about 1500.

3 The Empirical Benefits of Voluminous Data Sets

We provide two applications of extreme value theory to financial data, where the size of the data set is in some sense critical. First, we examine the properties of some foreign exchange (forex) contracts. It is shown how an increase in the data availability affects the tail estimates. Subsequently, we calculate hypothetical price returns on plain vanilla and exotic options. By using a large set of observations on the underlying we can break away from the parametric normality assumption, and study the implications of heavy tailed innovations on the derivative pricing process.

3.1 Forex Data Analysis

The first application focuses on the spot foreign exchange rate contract between the US Dollar and the German Mark (USD-DM). This contract is traded world wide over the telephone between a number of the larger banks. These banks maintain regularly updated buy and sell quotes at which they are willing to trade. These quotes are transmitted, via an information re-seller as Reuters, directly to the screens of currency traders. To give an idea about the size of the market, the daily turnover in the foreign exchange markets exceeds 10^{12} US dollars.

For a number of years the Olsen & Associates company has been collecting all the bid/ask quotes, and has made their data from October 1992 to September 1993 available to researchers. Over a 1.5 million quotes were given in the USD-DM contract during the year. In Daníelsson and de Vries (1997a) these data were transformed into returns by using the logarithmic difference in the average of the bids and asks quotes. Subsequently these data were used to calculate standardized quote returns on a ten minutes basis. From the ten minutes data base Daníelsson and de Vries (1998) and Daníelsson et. al. (1997) took a number of different subsamples. They calculated the Hill estimate of the tail index for these subsets. Their results are reported in Table 1. The tail index is between 3 and 4. This is fairly typical for financial return data. For most assets only the first few moments of the return distribution are bounded, a feature of the heavy tail property of the data. It implies, for example, that considerable care must be exercised in using statistics which have high moment requirements. In

Table 1: Tail Indexes for USD-DM spot contract

# Observations	$\widehat{1/\alpha}$	Gap
First 2000 10 minutes returns	0.10	0.25
Last 2000 10 minutes returns	0.35	
First 5000 10 minutes returns	0.30	0.07
Last 5000 10 minutes returns	0.37	
First 20000 10 minutes returns	0.27	0.03
Last 20000 10 minutes returns	0.30	
52000 all 10 minutes returns	0.25	

Estimates based on the Hill estimator (2) for the first and last subsample of the ten minutes aggregate returns and their differences. The number of order statistics is based on the subsample bootstrap technique from Daníelsson and de Vries (1998) the column labeled "Gap", the difference between the first and last subsample estimate is given. The value of increasing the sample size is shown by the decline in the Gap-size as the sample size increases.

3.2 Option Returns

We evaluate the return on an option contract written on the S&P500 stock index. The S&P500 index tracks the changes of an hypothetical portfolio of 500 different stocks listed on the New York Stock Exchange. The index is recalculated over the day as the underlying share prices change. Tradable options on the S&P500 are widely used in portfolio management.

Daníelsson and de Vries (1997b, 1998) examine the properties of the S&P500 index with special attention to the tail properties. The last 5000 daily returns from the daily S&P500 index, i.e. the period from 1984 to 1997, are used in the present analysis. Table 2 shows that a well diversified portfolio of American stocks yields an annualized return of about 10%, but this mean return carries quite a bit of uncertainty. The minimum daily return was -22%; this was the October 1987 Black Monday mini-crash. Combining the estimator (2) and (3) we calculate two quantile estimates. The borderline sample quantile estimate, i.e. the quantile that carries a probability of 1/5000, is less dramatic than the minimum referring to the Black Monday. The -8.62% and the -12.6% estimates show that -22.8% was a rare event indeed. Note that the estimate for the tail index is close to the estimates for the foreign exchange data.

Table 2: S&P500 return properties, Hill and quantile estimates

Mean	S.D.	Skewness	Kurtosis
10.0%	14.7%	-3.13	77.96
$\widehat{1/\alpha}$	Minimum	$\hat{X}_{1/n}$	$\hat{X}_{1/3n}$
0.35	-22.8	-8.62	-12.6
(0.31, 0.36)		(-7.82, -10.9)	(-11.4, -15.9)

The numbers between brackets give the 95% confidence band, with annualized percentages. Sample size $n = 5000$.

We use the S&P500 return data to construct the expected daily return on two options. A plain vanilla put option gives the owner the right to sell at a future point in time T the S&P500 index for a prefixed price X, where T is the *expiration date* or *maturity* and X is the *strike price*. At maturity the value of the put option, referred to as the *payoff*, is

$$\overline{PF} \equiv max(X - S_T, 0) \tag{9}$$

where S_T is the value of the S&P500 index at time T. Here T is equivalent to T^{TD}, the number of trading days. The present value of the payoff can be obtained through discounting

$$\overline{PV} = \overline{PF}\, exp(-r \times T^{CD}/365) \tag{10}$$

where r is the annual risk-free rate of interest and T^{CD} is the number of calendar days until expiration. We note that the number of trading days T^{TD} is usually less than the number of calendar days.

Because S_T is unknown, one needs to compute the expected value of $\overline{PV}$ for pricing the option. This can be done in at least two ways. The first method is based on the assumption that the S&P500 returns are normally distributed. This is the standard way of evaluating options in the financial literature, see e.g. Hull(1997). Under this assumption we estimate the mean and variance from the S&P500 return data, and subsequently generate pseudo normal random numbers. From these generated data we calculate the Monte Carlo expectation of $\overline{PF}$. This is merely a computer intensive way of evaluating the Black-Scholes formula (see Black and Scholes, 1973 and Boyle, 1977). In order to investigate the effect of non-normality, we generate bootstrap re-samples of the S&P500 returns vector and again compute the Monte Carlo expectation of $\overline{PF}$[1]. The assumption of normality is very convenient for two reasons. First, it permits one to express the value of the option in terms of elementary functions. Second, the method can proceed with few observations as only the mean and variance need to be estimated. A possible drawback is that one misprices the option because the frequency of the tail events is underestimated. Our procedure does not start from the assumption of normality, and therefore adequately captures the tail events. The drawback of this approach is the necessity for a large data set, because of the subsample bootstrap estimation procedure and the required presence of outliers. The sizeable S&P500 data set is sufficient to implement our procedure of fitting together the parametric and non-parametric parts of the distribution of the underlying.

Specifically, the bootstrap procedure is as follows. We create bootstrap one day S&P500 returns by re-sampling from the non-tail section of the empirical distribution and by drawing from the fitted tails outside the middle range. This procedure was proposed by Daníelsson and de Vries (1997b), where it is described extensively.

Let the one day return be denoted as y_{ij},with $i = 1, \ldots, T$, $j = 1, \ldots, N$, where N is the number of bootstraps. Hence the T day bootstrapped return is given by the sum of the single day returns

$$y_j = \sum_{i=1}^{T} y_{ij} \tag{11}$$

Conditional on one simulated T-period-return y_j, the simulated value of the S&P500 at the expiration date T is

$$S_T = S_0\, exp(r \times T/365) exp(y_j)/(\bar{y}^T) \tag{12}$$

[1] The validity of this approach hinges on the assumption that the market prices the derivatives as if agents were risk neutral, or alternatively that the derivatives can be priced by a dynamic risk neutral hedge strategy. Our procedure is as in Boyle (1977), except that we also allow for non-normality.

and

$$\bar{y} = \left(\frac{1}{n} \sum_{k=2}^{n} \frac{P_k}{P_{k-1}} \right) \tag{13}$$

where $T = T^{CD}$, n is the sample window size, P_k is the S&P500 index price on day k and S_0 is the current value of the S&P500 index at the beginning of the option life time, see Boyle (1977). By construction the factor $exp(y_j)/\bar{y}$ has bootstrap expectation $E[exp(y_j)]/(\bar{y}^T) = 1$. Thus the factor generates the distribution of S_T centered around the futures price.

The present value of the expected option payoff is then estimated by averaging over the number of bootstraps

$$\overline{OP} \equiv \frac{1}{N} \sum_{j=1}^{N} \overline{PV} \tag{14}$$

Denote the expected present value of the option as of tomorrow by $\overline{OP}^*$. To obtain $\overline{OP}^*$, first update S_0 by a randomly drawn single day return y_{ij}. Subsequently, repeat the above procedure for the time to maturity $T-1$. Then the one day option return is

$$y^{OP} = \ln \left(\frac{\overline{OP}^*}{\overline{OP}} \right) \tag{15}$$

Apart from this put option, we also investigate the return on an exotic option. Specifically, we considered an up-and-out put barrier option, with a discrete barrier $H > X$, such that if at any time $\tau \in (1,T)$ we have that $S_0\, exp(r \times \tau/365)\, exp(y_{\tau j})/(\bar{y}^\tau) \geq H$ with $y_{\tau j} = \sum_{i=1}^{\tau}(y_{ij})$, the option expires and its payoff is zero. Otherwise its payoff is defined in the same way as for the put option.

The two options were evaluated by using $S_0 = 930.87$, the S&P500 index value on September 4, 1997. The strike price X was set at 950. The number of calendar days to maturity was $T^{CD} = 105$, while the number of trading days to maturity was $T^{TD} = 76$. We took as the annual risk-free interest rate the 3 months US Treasury-Bill rate, which at that time was 5.5%. The number of bootstraps is indicated by N. We also limited ourselves to consider only the last $n = 1500$ observations of the S&P500 data set, from September 30, 1991 to September 4, 1997 (where $P_1 = 387.86$), in order to strike a balance between frequent updating, as is commonplace in option pricing, and the data demands by the bootstrap estimation procedure. The 1500 is about the minimum number of observations we need for the bootstrap procedure. We considered two possible barrier levels for the exotic option. One was $H = (1 + 20\%)X$ and the other was $H = (1 + 15\%)X$. We also evaluated the options by using a data set of normally distributed random variables with the same mean and standard error of the S&P500 data set.

The results of our simulations are collected in Table 3. The Table 3 gives the

Table 3: One day options' returns

Options	Normally Distributed Data	S&P500 Data
Put	0.0116	0.0179
	(0.0034)	(0.0034)
Barrier 20%	0.0086	0.0162
	(0.0034)	(0.0034)
BB	175,409	158,249
Barrier 15%	0.0092	0.0169
	(0.0034)	(0.0034)
BB	1,668,616	1,504,574

We used N=10^5 with the total number of calculation equal to $15 \cdot 10^7$ for each simulation. The table gives the one day return on the options. The numbers between brackets are the standard errors of the mean of the one day option returns, and BB indicates the number of breaches of the barrier.

bootstrap average of the daily return y^{OP}, the standard error of the mean and the number of times the barrier was breached. From Table 3 we see that the one day option returns are all quite similar, which is not so surprising given the time scale of a single day. The returns based on the normally distributed data set are, however, somewhat below the S&P500 based returns. This may be due to the difference in the tail shapes, but at this stage we have no good explanation for this phenomena. We plan to investigate this result further in future research. The Table 3 also shows that the introduction of a barrier has no affect on the one-day option returns which is due to the imposed risk neutrality. But, there is a difference in how often the barrier is breached. As expected, the closer the barrier is to the exercise price, the more frequently the barrier is crossed. Moreover, the number of times the barrier is crossed is higher for the normal based data than for the S&P500 based returns. The reason for this apparently counter-intuitive result follows from the properties of extremes. Recall that S_T is calculated by using the T-day return y_j on the S&P500. This T-day return is obtained by summing T one day returns y_{ij}. Under normality a convolution of T i.i.d. variables enlarges the scale of the single day return by a factor $\sqrt{T}$. For heavy tailed distributions we can rely on Feller's (1971, VIII.8) theorem on the tail shape of a convolution of i.i.d. heavy tailed variables. If the single heavy tailed variable has the tail probability

$$\mathrm{P}\left\{Y > q\right\} = aq^{-\alpha} \tag{16}$$

then for the T times convolution

$$\mathrm{P}\left\{\sum_{i=1}^{T} Y_i > q\right\} = T\,aq^{-\alpha} \tag{17}$$

Hence, in the tails the scale of the quantiles is increased by a factor $T^{1/\alpha}$ if we hold the probability level constant. Now note that we find $\alpha > 2$, and hence the scale increase is less pronounced if the return data exhibit a heavy tail than if these were normally distributed. Thus, the stock price is more likely to be above the strike price at the expiration date, or will more likely hit the barrier H, under normality than if the returns were heavy tailed. Hence, for both barrier options, the normal based put options hit the barrier 10% more often than the heavy tail based one. It may sound counter intuitive that the heavy tailed model has less chance of crossing the barrier, since the tails are heavy. If the barrier were very high, one would find indeed that the heavy tailed data do have a higher chance of hitting the barrier. But for more moderate values, the aggregation effect dominates over the hyperbolic tail shape effect.

4 Conclusions

There are at least two advantages in using large financial data sets. First, such data sets enable a precise study of the risks on extreme losses (and gains) through semi-parametric tail estimation. In an application to exotic options pricing, we showed how the tail based method uncovered that the normal based pricing technique has a bias to hitting the barrier. Without a sizeable data set this bias would have been hard to detect. We also found that the options returns from the normal based data set are somewhat below the returns that are computed with the heavy tailed S&P500 data set. Whether this is a robust result that can be explained on the basis of the tail shape differences, requires further scrutiny. Second, if the size of the data set can be enlarged, then the AMSE of the tail procedure is enhanced. We illustrated this by estimating the tail index of a foreign exchange rate return for different sample sizes. In the future we plan to extend the option pricing application to a high-frequency data set as well.

References

[1] Baillie, R. and P. McMahon (1989). *The Foreign Exchange Market, Theory and Econometric Evidence*, Cambridge University Press, Cambridge.

[2] Black, F. and M. Scholes (1973). The pricing of options and corporate liabilities. *Journal of Political Economy, 81, 637-659.*

[3] Boyle, P.P. (1977). Options: A Monte Carlo approach. *Journal of Financial Economics 4, 323-338.*

[4] Campbell, J.Y., A.W. Lo and A.C. MacKinlay (1997). *The Econometrics of Financial Markets*, Princeton University Press, Princeton.

[5] Dacorogna, M.M., U.A. Muller, O.V. Pictet and C.G. de Vries (1995; revised 1998). The distribution of extremal foreign exchange rate returns in extremely large data sets. *Tinbergen Institute Discussion Paper, TI 95-70.*

[6] Daníelsson,J., L. de Haan, L. Peng and C.G. de Vries (1997). Using a bootstrap method to choose the sample fraction in tail index estimation. *Tinbergen Institute Discussion Paper, TI 97-016/4.*

[7] Daníelsson, J. and C.G. de Vries (1997a). Tail index and quantile estimation with very high frequency data. *Journal of Empirical Finance 4, 241-257.*

[8] Daníelsson, J. and C.G. de Vries (1997b). Value-at-Risk and extreme returns. *London School of Economics FMG Discussion Paper No 273.*

[9] Daníelsson, J. and C.G. de Vries (1998). Beyond the sample: Extreme quantile and probability estimation. *Tinbergen Institute Discussion Paper, TI 98-016/2.*

[10] De Haan, L. (1990). Fighting the arch-enemy with mathematics. *Statistica Neerlandica, 44, 45-68.*

[11] De Haan, L., D.W. Jansen, K. Koedijk and C.G. de Vries (1994). Safety first portfolio selection, extreme value theory and long run asset risks. *In J. Galambos, J. Lechner, E. Simiu and N. Macri (eds.), Extreme Value Theory and Applications, 471-487.*

[12] De Haan, L., S.I. Resnick, H. Rootzen and C.G. de Vries (1989). Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes . *Stochastic Processes and their Applications, 32, 213-224.*

[13] De Haan, L. and U. Stadtmüller (1996). Generalized regular variation of second order. *Journal of the Australian Mathematical Society A 61, 381-395.*

[14] Engle, R.F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of the UK inflation. *Econometrica 50, 987-1008.*

[15] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol.II 2nd ed. Wiley, New York.

[16] Goldie, C.M. and R.L. Smith (1987). Slow variation with remainder: Theory and applications. *Quarterly Journal of Mathematics 38, 45-71, Oxford 2nd series.*

[17] Hall, P. (1982). On some simple estimates of an exponent of regular variation. *Journal of Royal Statistical Society B 42, 37-42.*

[18] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics 35, 1163-1173.*

[19] Hull, J. (1997). *Options, future, and other derivative securities.* Prentice-Hall International.

[20] Kearns, P. and A. Pagan (1998). Estimating the density tail index for financial time series. *Review of Economics and Statistics 79, 171-175.*

[21] Resnick, S. and C. Starica (1998). Tail estimation for dependent data. *Annals of Applied Probability, forthcoming.*