COMISEF
Computational Optimization Methods in Statistics, Econometrics and Finance

- Marie Curie Research and Training Network funded by the EU Commission through MRTN-CT-2006-034270 -

COMISEF WORKING PAPERS SERIES

WPS-006 08/10/2008

Least Median of Squares Estimation by Optimization Heuristics with an Application to the CAPM and Multi Factor Models

P. Winker
M. Lyra
C. Sharpe

Least Median of Squares Estimation by Optimization Heuristics with an Application to the CAPM and Multi Factor Models*

Peter Winker, Marianna Lyra and Chris Sharpe
Department of Economics, Justus-Liebig University Giessen
{Peter.Winker,Marianna.Lyra,Chris.Sharpe}@wirtschaft.uni-giessen.de

August 4, 2008

Abstract

For estimating the parameters of models for financial market data, the use of robust techniques is of particular interest. Conditional forecasts, based on the capital asset pricing model, and a factor model are considered. It is proposed to consider least median of squares estimators as one possible alternative to ordinary least squares. Given the complexity of the objective function for the least median of squares estimator, the estimates are obtained by means of optimization heuristics. The performance of two heuristics is compared, namely differential evolution and threshold accepting. It is shown that these methods are well suited to obtain least median of squares estimators for real world problems. Furthermore, it is analyzed to what extent parameter estimates and conditional forecasts differ between the two estimators. The empirical analysis considers daily and monthly data on some stocks from the Dow Jones Industrial Average Index (DJIA).

Keywords: LMS, CAPM, Multi Factor Model, Differential Evolution, Threshold Accepting.

*Financial support from the EU Commission through MRTN-CT-2006-034270 COMISEF is gratefully acknowledged.

1 Introduction

Despite of its attractive theoretical features, the estimation and analysis of the capital asset pricing model (CAPM) and other models with more than one factor is often complicated by the fact that the distribution of the error terms cannot be assumed to be independently identically normal. Consequently, different robust estimation approaches have been considered (see, e.g., Chan and Lakonishok (1992), Knez and Ready (1997), Martin and Simin (2003), Ronchetti and Genton (2008)). In this contribution, we consider the classical least median of squares (LMS) estimator (Rousseeuw 1984).

Although this estimator exhibits nice properties with regard to robustness, it is not used frequently. One possible reason is that estimation requires to solve a complex optimization problem. In particular, the objective function landscape is not smooth and exhibits many local optima. Consequently, traditional optimization methods will fail. One alternative consists in exploiting the inherent discrete nature of the optimization problem and resorting to a full enumeration of all potential solutions. An algorithm built on this approach is PROGRESS proposed by Rousseeuw and Leroy (1987).[1] However, the complexity of this approach grows at a rate of T^2 in the sample size T for the bivariate regression. If more than one factor has to be considered, the complexity becomes even worse. Furthermore, the technique does not allow for a simple implementation of nonlinear or constraint estimation. These shortcomings might be overcome by optimization heuristics.

Heuristic optimization techniques have been successfully applied to a variety of problems in statistics and economics for well over a decade (see Gilli *et al.* (2008) and Gilli and Winker (2008) for recent overviews). However, applications to estimation problems are still rare. Fitzenberger and Winker (2007) consider Threshold Accepting (TA) for censored quantile regression, a problem similar to the LMS estimator.[2] Maringer and Meyer (2008) and Yang *et al.* (2007) also use TA for model selection and estimation of smooth transition autoregressive models. In contrast, several optimization heuristics have been used in other fields of research in finance, e.g., portfolio optimization (Dueck and Winker (1992), Maringer (2005), Winker and Maringer (2007*a*), Specht and Winker (2008)) or credit risk bucketing (Krink *et al.* 2007).

We present an application to the LMS estimator. In particular, we propose implementations of Threshold Accepting (TA) and Differential Evolu-

[1] Barreto and Maharry (2006) propose a generalization for the bivariate regression without a constant. The approach might be considered as an application of elemental subset regression (Mayo and Gray 1997).

[2] In fact, Fitzenberger and Winker (2007) exploit the elemental subset properties of quantile regression for their approach.

tion (DE) for obtaining the LMS estimator. We purposely select a population based search method, DE, and a local search method, TA, to compare their efficacy on a continuous search space.[3] We provide some evidence on the tuning of both algorithms and the relative performance for this problem. It turns out that the LMS estimator can be obtained quite reliably using optimization heuristics despite of its high inherent complexity.

Finally, we apply the estimator to the CAPM and a three factor model for a large set of rolling window samples for some of the stocks comprising the Dow Jones Industrial Average Index (DJIA). The estimates differ substantially for some stocks and time periods from those obtained by ordinary least squares (OLS). We also calculate the conditional forecasts based on the model and the actual factor values. These conditional forecasts are compared with those obtained from the OLS estimates. It is also analyzed to what extend a combination of both forecasts might reduce the forecasting errors.

The rest of the study is organized as follows. Section 2 shortly reviews the theoretical background to the underlying models of the financial market and the LMS estimator. Section 3 reports on heuristic strategies, describes the optimization problem and the algorithms used. The specific application and the empirical results are presented in Section 4. In Section 5, we provide evidence on the rate of convergence of the two heuristics, while Section 6 summarizes the main findings and provides an outlook to further research.

2 Theoretical Background

2.1 CAPM and Multi Factor Models

Traditionally, the capital asset pricing model (CAPM) provides the method for estimating the risk-return equilibrium. The pioneering work by Markowitz (1952) has set the foundation of modern portfolio management and was employed later by Sharpe (1964), Lintner (1965) and Mossin (1966) to develop the CAPM. The CAPM describes a linear relationship between the risk premium on individual securities relative to the risk premium on the market portfolio. It is given by

$$r_{i,t} - r_t^s = \alpha + \beta(r_{m,t} - r_t^s) + \varepsilon_{i,t}\,, \tag{1}$$

where

[3]Note, that we do not take into account the implicit discrete structure of the optimization problem related to elemental subset regression. This might reduce the performance of TA, but allows to introduce constraints and nonlinear components in future research.

$r_{i,t}$	rate of return at time t for asset i
r_t^s	risk free rate of return at time t
α, β	parameters of CAPM
$r_{m,t}$	market rate of return at time t
$\varepsilon_{i,t}$	residual at time t for asset i .

The simplicity of the CAPM, i.e., the concentration on a single risk factor, is one of the reasons why the explanatory power of the model is limited. One extension to the model has been proposed by Fama and French (1992). The authors emphasize the multi-dimensionality of risks. In particular, they propose to consider the effects of firm size and book value to equity in explaining the cross-section of average stock returns. In another paper, Fama and French (1993) introduce the three-factor model. They conclude that the market factor together with a size and a book-to-market factor can explain 95% of the variation in excess stock returns. A key finding is that the difference between small and big firms and the difference between high and low value captures variation through time.[4]

The final form of the model used in our application is given by equation (2). While the market risk $r_{m,t} - r_t^s$ is as for the CAPM given by the difference between the market rate of return and the risk free rate, *SMB* is defined as the average returns on three portfolios comprising small firm stocks minus the average return on three portfolios comprising larger firms, and *HML* is the average return on two portfolios comprising firms with high book-to-market value minus the average return on the two so called growth portfolios with low book-to-market values.[5]

$$r_{i,t} - r_t^s = \alpha + \beta_1(r_{m,t} - r_t^s) + \beta_2 SMB_t + \beta_3 HML_t + \varepsilon_{i,t}\,, \tag{2}$$

where

$r_{m,t} - r_t^s$	factor accounting for market risk premium at time (t)
SMB_t	factor accounting for size premium at time t
HML_t	factor accounting for value premium at time t
$\beta_{1,2,3}$	exposure levels to the corresponding risk factors .

[4]For a critical assessment of the empirical performance of the model, see, e.g., Knez and Ready (1997).

[5]In the empirical application, we will use data for these factors provided by the authors on their website.

2.2 LMS

A substantial amount of research in financial market economics has focused on estimating the parameters of CAPM and multi factor models.[6] OLS estimation can be problematic due to its lack of robustness (Rousseeuw and Wagner (1994), Ronchetti and Genton (2008)). In particular, outliers can have a strong effect on the estimated coefficients.[7] The smallest percentage of influential observations that can change the parameters of the regression line is called breakdown point (Rousseeuw 1984).

In order to achieve a higher breakdown point, a number of robust techniques have been suggested in the statistical literature. These include least absolute deviations (LAD), also called minimum absolute deviations (MAD) suggested by Sharpe (1971), Cornell and Dietrich (1978), and later by Chan and Lakonishok (1992). The former group of authors applied also trimmed regression quartile (LTQ) estimators. Further, least trimmed squares (LTS) were proposed by Zaman *et al.* (2001) and more recently M-estimators by Martin and Simin (2003). In the present application we concentrate on the least median of squares (LMS) estimator, introduced by Rousseeuw and Leroy (1987).

The main purpose of the implementation is not to demonstrate the superior performance of LMS estimates, e.g., with regard to predictive performance,[8] but to provide a proof of concept, i.e., that LMS estimates obtained by means of heuristic optimization can be used for real life applications. This will allow for further extensions in the future, e.g., taking into account constraints or nonlinear relationships.

The LMS estimator for the CAPM is defined as the solution to the following optimization problem:

$$\min_{\alpha,\beta}(\text{med}(\varepsilon_{i,t}^2)) \,, \tag{3}$$

where $\varepsilon_{i,t} = (r_{i,t} - r_t^s) - \alpha - \beta(r_{m,t} - r_t^s)$ according to equation (1) above. This results in a highly complex objective function as exhibited for one problem instance in Figure 1.

For the multi factor model (2), the optimization problem becomes,

[6] We will not comment on the difficulties related to the definition of the variables, in particular the risk free rate of return and – even more difficult – the market rate of return which should summarize all available investment opportunities.

[7] In the application to financial market data, the meaning of "outliers" is not obvious. In fact, they might provide relevant information and should not be discarded from the analysis (Knez and Ready 1997).

[8] For a discussion of the predictive performance of CAPM and multi factor models based on OLS estimates see Simin (2008).

$$\min_{\alpha,\beta_1,\beta_2,\beta_3} \left(\text{med}(\varepsilon_{i,t}^2)\right), \tag{4}$$

with $\varepsilon_{i,t} = (r_{i,t} - r_t^s) - \alpha - \beta_1(r_{m,t} - r_t^s) - \beta_2 SMB_t - \beta_3 HML_t$.

3 Heuristic Strategies

Recent research[9] suggests that even an apparently simple optimization problem might result in an objective function which does not allow for the successful application of standard numerical approximation algorithms. In this vein, minimizing the median of squared residuals results in a search space containing many local minima, where traditional optimization methods can not provide an exact solution. As an example, Figure 1 shows the above objective function using the 200 daily stock returns of the IBM stock starting on January, 2nd, 1970.[10]

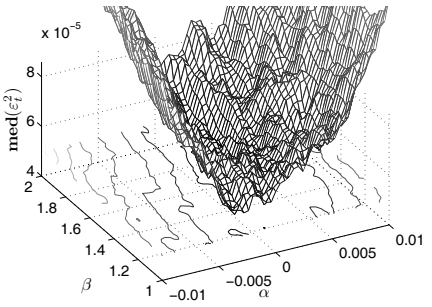

Figure 1: Median of squared residuals as a function of α and β.

In principle, it is possible to provide an exact solution to this optimization problem by exploiting the inherent discrete structure of the problem. However, this comes at high computational cost which becomes a binding constraint when additional factors are considered. Furthermore, the technique based on elemental subset regression proposed by Rousseeuw and Leroy (1987) can not easily be generalized for nonlinear models or when some constraints are imposed on the parameter space. Alternatively, heuristic optimization methods are well suited to handle such problems. If traditional

[9] E.g., Gilli and Winker (2007) and the papers in that special issue.

[10] In passing note that the sample size used for the estimation in our application is less than one year. Thus, it is substantially lower than the standard practice of 5 years used in industry (Simin 2008, p. 358). However, repeating our analysis with larger samples does not affect the qualitative findings.

methods fail due to the existence of many local optima, the performance of optimization heuristics will typically dominate them in terms of solution quality. In the following, we will analyze the performance of two heuristic methods for the LMS estimation problem, TA and DE.

3.1 Threshold Accepting

Originally devised by Dueck and Scheuer (1990), Threshold Accepting (TA) has proven to be a simple, powerful search tool for many types of optimization problem. A key advantage of TA is that it enables the search to escape local minima. Here we present a modified version of the standard TA algorithm for the LMS estimation problem.[11] Algorithm 1 provides the general outline. First, the number of rounds n_R and the number of steps per round n_S are initialized as well as the threshold sequence τ_r. Next, a random solution x^0 is chosen (2:). Then, for each round, n_S local search steps are executed for a fixed value of the threshold, which determines (6:) to what extent not only local improvements, but also local impairments are accepted. The algorithm terminates after the a priori fixed number of $n_R \times n_S$ iterations.

Algorithm 1 Threshold Accepting Algorithm.

```
 1: Initialise n_R, n_S , and τ_r, r = 1, 2,...,n_R
 2: Generate at random a solution x^0 ∈ [α_l α_u] × [β_l β_u]
 3: for r = 1 to n_R do
 4:    for i = 1 to n_S do
 5:       Generate neighbour at random, x^1 ∈ N(x^0)
 6:       if f(x^0) − f(x^1) < τ_r then
 7:          x^0 = x^1
 8:       end if
 9:    end for
10: end for
```

We shall first look at the generation of the threshold sequence τ_r, which has a fundamental effect on the search behavior of the algorithm. Broadly speaking, the TA search behaves like a random search in the initial stages, for large values of τ, and gradually transforms into a greedy search, as $\tau \to 0$. The degree of randomness in the initial stages depends on the starting value of the threshold sequence; the degree of 'greediness' in the latter stages depends on how the threshold is reduced. Rather than guess the two extents of the threshold sequence and refine the values by trial and error, Winker and Fang (1997) suggested a data driven approach (see also Winker and Maringer

[11] A description of the general form of the algorithm and its behavior is given by Winker and Maringer (2007*b*). For a further, comprehensive overview see Winker (2001).

(2007*b*)). The pseudocode for the data driven generation of the threshold sequence is provided in Algorithm 2.

Algorithm 2 Data Driven Generation of Threshold Sequence.

1: Initialize n_R, lower quantile α, $n_D = [n_R/\alpha]$
2: **for** $r = 1$ to n_D **do**
3: Generate at random a solution $x_r^c \in [\alpha_l\ \alpha_u] \times [\beta_l\ \beta_u]$
4: Generate at random a near neighbour solution $x_r^n \in \mathcal{N}(x_r^c)$
5: Calculate $\Delta_r = |f(x_r^n) - f(x_r^c)|$
6: **end for**
7: Sort $\Delta_1 \leqslant \Delta_2 \leqslant \cdots \leqslant \Delta_{n_D}$
8: Use $\Delta_{n_R}, \ldots, \Delta_1$ as threshold sequence

Returning to the main TA algorithm, we select reasonable boundaries for the search space $[\alpha_l \alpha_u] \times [\beta_l \beta_u]$ and generate a starting point x^0 at random within this area (2:). Then, in each round r, a further n_S solutions are randomly generated within a neighborhood of the current solution, $N(x^0)$, and for each solution the objective function is computed, subtracted from the current solution, $f(x^0)$, and adopted as the new solution if the result of the calculation is less than the threshold.

There are several options for the shape of the neighborhood in a two dimensional search space. However, a hyper-rectangle offers the advantage of small computational overhead to recalculate its dimensions.[12] We set the initial dimensions of the hyper-rectangle to the boundaries of the search space and reduce the dimensions proportionally with the number of rounds. Whether we choose a linear reduction or geometric reduction of the neighborhood dimensions does not appear to make much of a difference for the quality of the results. The number of reductions to the hyper-rectangle is n_R, equal to the number of values in the threshold sequence. The rationale behind reducing the neighborhood is closely connected to the behavior of the search and the threshold sequence. The first neighborhood allows for a general exploration of the full search space; at the end there is a limited, concentrated exploration of a small area, where we assume good quality solutions lie.

One issue that needs to be dealt with is reconstructing the neighborhood if it exceeds the bounded search space. A straight forward solution is to shift the whole neighborhood in a vertical, horizontal, or diagonal direction by the amount it has exceeded the bounded search space. It should be emphasized that this would be a less trivial, computationally more expensive operation with other neighborhood shapes.

[12] See Winker (2001) and Gilli *et al.* (2008) for a general discussion on neighborhoods in higher dimensional search spaces.

3.2 Differential Evolution

DE is a population based optimization technique for continuous objective functions developed by Storn and Price (1997). The algorithm starts with a randomly initialized set of candidate solutions. Then, for a predefined number of generations, the elements of the population are updated by generating linear combinations of existing elements and random crossover. Finally, the objective function value of the new candidate solution is compared with that of the original element. If it is lower, the new candidate solution replaces the old one. Algorithm 3 provides the pseudocode of our implementation.

Algorithm 3 Differential Evolution.

```
 1: Initialize parameters n_p, n_G, F and CR
 2: Initialize population P^(1)_{j,i}, j = 1, ···, d, i = 1, ···, n_p
 3: for k = 1 to n_G do
 4:    P^(0) = P^(1)
 5:    for i = 1 to n_p do
 6:       Generate r1,r2,r3 ∈1, ···,n_p, r1 ≠ r2 ≠ r3 ≠ i
 7:       Compute P^(v)_{.,i} = P^(0)_{.,r1} + F × (P^(0)_{.,r2} - P^(0)_{.,r3})
 8:       for i = 1 to d do
 9:          if u < CR then
10:             P^(u)_{j,i} = P^(v)_{j,i}
11:          else
12:             P^(u)_{j,i} = P^(0)_{j,i}
13:          end if
14:       end for
15:       if f(P^(u)_{.,i}) < f(P^(0)_{.,i}) then
16:          P^(1)_{.,i} = P^(u)_{.,i}
17:       else
18:          P^(1)_{.,i} = P^(0)_{.,i}
19:       end if
20:    end for
21: end for
```

As mentioned above, the initial population of n_p elements is randomly chosen (2:). Then, for a predefined number of generations n_G, the algorithm performs the following procedure. Each element of the population is updated by means of differential mutation (7:) and crossover (9:). Particularly, differential mutation constructs new parameter vectors by adding the scaled difference of two randomly selected vectors to a third one. F is the scale factor that determines the speed of shrinkage in exploring the search space. During crossover, DE recombines the initial elements with the new candidates by replacing each component $P^{(0)}_{j,i}$ with a probability of CR with mutant ones $P^{(v)}_{j,i}$ resulting in a new trial vector $P^{(u)}_{j,i}$. Finally, the value of

the objective function of the trial vector is compared with that of the initial element. Only if the trial vector results in a better value of the objective function, it replaces the initial element in the population. The above process repeats until all elements of the population have been considered. Then, the process restarts for the next generation.

Calibration Issues

Price *et al.* (2005) report that, although the scale factor F has no upper limit and the crossover parameter CR is a fine tuning element, both are problem specific. In an attempt to improve the tuning of the algorithm, we conducted repeated runs for different values of the population size n_p and the number of generations n_G. During this initial phase we did not tune the weighting (scaling) factor (F) and the crossover probability (CR). In order to achieve convergence, we increased the population size n_p to more than ten times the number of parameters.[13] We observe that when the best value is found repeatedly for several runs of the algorithm, a further increase in the number of generations (to more than 100) does not improve the results, while the computational time increases. With a population size of $n_p = 20$, which is ten times the number of parameters for the CAPM (2), a number of generations of $n_G = 100$, and the constants set to $F = 0.8$ and $CR = 0.9$, the algorithm typically converges to the same results in several replications. By increasing the population size to $n_p = 50$, the algorithm consistently provides identical outcomes in each repetition.

For fine tuning the technical parameter, the algorithm has been run for different combinations of F and CR. The procedure is illustrated in Algorithm 4 for parameter values ranging from 0.5 to 0.9.

Algorithm 4 Calibration Issues.

1: Initialize parameters n_p, n_G
2: Initialize population $P_{j,i}^{(1)}$, $j = 1, \cdots$, d, i $= 1, \cdots, n_p$
3: **for** $F = 0.5, \cdots, 0.9$ **do**
4: **for** $CR = 0.5, \cdots, 0.9$ **do**
5: Repeat Algorithm 3 from line 3-21
6: **end for**
7: **end for**

Figure 2 exhibits the dependence of the best value of the objective function obtained for different combinations of F and CR always for the same problem instance (first 200 observations of the IBM stock in our sample).

[13] A practical advice for optimizing objective functions with DE given on www.icsi.berkeley.edu/∼storn/.

The population size n_p and the number of generations n_G are set to 50 and 100, respectively. The left side of Figure 2 presents the results for a single run of the algorithm, while the right side shows the mean over 30 restarts. Although the surface is full of local minima for CR below 0.7, it becomes smoother as CR reaches 0.8 independent of the choice of F. The results clearly indicate that for higher values of CR, results improve, while the dependency on F appears to be less pronounced. Based on these results, we use $F = 0.8$ and $CR = 0.9$ for estimating the parameters of the models in the next section.

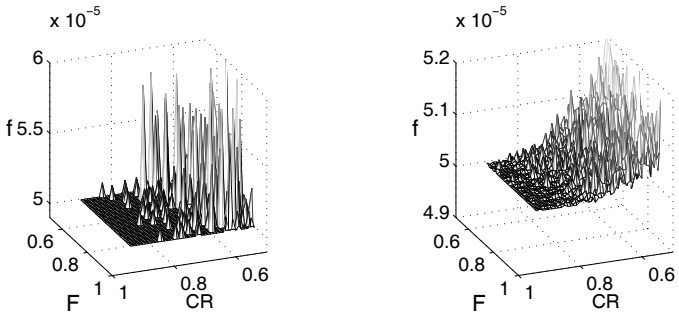

Figure 2: Calibration of technical parameters.

4 Empirical Findings

4.1 Implementation Details

For the application of LMS to the CAPM, we consider daily data from the sample of publicly traded firms comprising the Dow Jones Industrial Average for the period between 1970 and 2006. We select six companies, IBM, ExxonMobil (XOM), General Electric (GE), Merck (MRK), General Motors (GM) and Boeing (BA).

In order to estimate the parameters of the CAPM over a sensible time period, we use a rolling window of 200 days length moving from 1970 to 2006 day by day. For each given sample, the parameters α and β are obtained by LMS estimation using ten restarts of each heuristic.[14] The estimates corresponding to the best value of the objective function are kept.

[14]We only report results for the DE implementation as it appears to be more efficient for the specific problem as discussed in Section 5 below. For the DE implementation, we use $n_P = 50$ and $n_G = 100$. The computation time for 10 restarts on a given sample amounts to about 1.7 seconds using Matlab 7.4 on a PC with Intel Duo Core processor operating at 2.39 GHz, running Windows XP OS.

Next, for given parameter estimates, we calculate the forecast of the excess return conditional on the market return for the next trading day. The same calculation is done based on the OLS estimates. Finally, both forecast errors are compared. Algorithm 5 summarizes the procedure.

Algorithm 5 Rolling window estimation.

1: Initialize parameters
2: **for** $t = 1$ to 9113 **do**
3: Run optimization heuristic 10 times for sample $[t \dots t + 199]$
4: LMS estimates α_t^{LMS} and β_t^{LMS} correspond to best value of objective function
5: OLS estimates α_t^{OLS} and β_t^{OLS}
6: Calculate one-day-ahead conditional forecasts
7: **end for**

For the Fama/French multi factor model we use monthly data for the period between 1962 and 2008, except for XOM and MRK stocks for which the sample starts only in 1970.[15] The length of the rolling window is fixed to 10 months. Otherwise, the procedure is identical to that used for the CAPM.

4.2 Estimation Results

The results of the rolling windows estimation for the CAPM are illustrated in Figures 3 and 4, for IBM and XOM, respectively. In both figures the actual stock returns are presented in the top graph, for the period between 1970 and 2006. The β estimates using OLS and LMS are presented in the middle and the bottom graphs, respectively.[16] We include the stock returns graph in order to identify whether LMS performs better in periods where larger fluctuations in stock prices are observed. Despite the expectations that LMS estimators will be smoother, since by definition they are not influenced by extreme observations, the opposite result seems to occur not only for IBM and XOM but also for the other stocks in our analysis.

Obviously, the results should not be taken as general findings on the relative performance of OLS and LMS. In particular, we have to study the following issues in more detail: the effect of sample size; the volatility of stock returns in the period considered; the adequacy of the model.

[15] The data for the factors are taken from Kenneth R. French's website `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

[16] For a few samples we compared our results for the CAPM with those obtained from the `R` implementation of LMS which allows to derive exact solutions by full enumeration. Typically, the estimates obtained by DE are almost identical to those results. However, for a few cases, we found slightly better values for the estimates from DE which might point at some numerical problems with the `R` implementation.

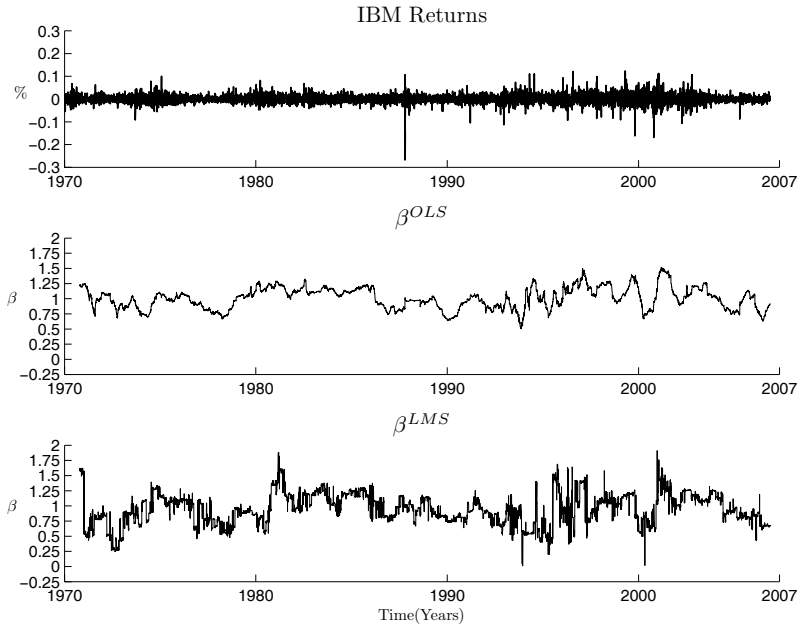

Figure 3: Estimates of β for IBM and the period between 1971 and 2006.

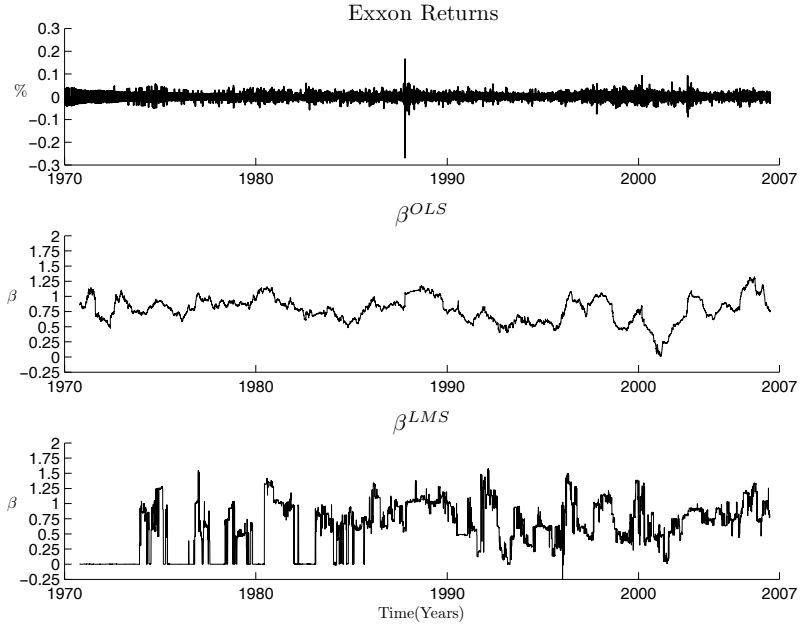

Figure 4: Estimates of β for Exxon and the period between 1971 and 2006.

First, the effect of a single outlier in the OLS estimates will be more pronounced for smaller samples. However, for financial market data, the issue might be less one of singular cases, but rather of fat tails, i.e., frequent extreme observations. Nevertheless, in future research we plan to repeat the analysis for different sample lengths. Second, we considered the stocks of two large companies for a long time period. It is certainly worth considering alternative stocks, e.g., high-tech stocks with small capitalization, and to identify sub-periods for which the relative performance of OLS and LMS differ most. Third, we also considered the three factor model proposed by Fama and French (1992). The estimation results are available on request. We will refer to this model with regard to its forecasting performance in the next subsection.

4.3 Forecasting Performance

Although the estimates of the CAPM and the multi factor model might be of interest on their own, the typical application consists in using them for (conditional) forecasting (Simin 2008). Given the marked differences between LMS and OLS estimates for both models, it is of interest to see how this affects the predictive performance. To this end, we calculated the mean squared forecast error (MSE) and the mean absolute forecast error (MAE) for the one-period-ahead conditional forecasts for 9113 days for the CAPM and up to 555 months for the multi factor model. Table 1 summarizes the findings for the CAPM.

Is is remarkable that typically the forecasts based on OLS estimates do not only exhibit smaller MSE, but also smaller MAE. While the first finding might have been expected given that in sample MSE is the objective function for OLS, for MAE the robustness of the LMS estimator could have led to a different result. The only exception is the XOM stock, for which the MSE can be reduced substantially when using LMS instead of OLS.

Table 2 reports the MSE and the MAE for the multi factor model.[17]

So far, we might see our results as a further contribution to the rather disappointing evidence regarding the predictive performance of factor models (Simin 2008). In particular, no improvement over conventional OLS based forecasts is apparent. However, this findings do not exclude that the LMS based forecasts outperform the OLS approach at least under specific market conditions, e.g., high versus low volatility regimes. An analysis of this aspect is left to future research.

[17]When using a rolling window of length 25 months the difference between MSE and the MAE using LMS and OLS become smaller, but still the predictive performance of OLS appears to be superior.

Table 1: Forecast errors for LMS and OLS estimates of the CAPM.

Stock		LMS	OLS
IBM	MSE	$0.1835 \cdot 10^{-3}$	$0.1784 \cdot 10^{-3}$
	MAE	0.0092	0.0090
XOM	MSE	$0.1628 \cdot 10^{-3}$	$0.1882 \cdot 10^{-3}$
	MAE	0.0092	0.0093
GE	MSE	$0.1760 \cdot 10^{-3}$	$0.1409 \cdot 10^{-3}$
	MAE	0.0090	0.0087
MRK	MSE	$0.2236 \cdot 10^{-3}$	$0.1974 \cdot 10^{-3}$
	MAE	0.0107	0.0099
GM	MSE	$0.2233 \cdot 10^{-3}$	$0.2115 \cdot 10^{-3}$
	MAE	0.0106	0.0103
BA	MSE	$0.3903 \cdot 10^{-3}$	$0.3471 \cdot 10^{-3}$
	MAE	0.0136	0.0133

Table 2: Forecast errors for LMS and OLS estimates of multi factor model.

Stock		LMS	OLS
IBM	MSE	26.3819	9.7380
	MAE	3.3231	2.3399
XOM	MSE	11.7523	4.8583
	MAE	2.4383	1.6077
GE	MSE	16.6303	6.1978
	MAE	2.9205	1.9088
MRK	MSE	24.8864	21.9948
	MAE	3.4414	3.5744
GM	MSE	24.0990	17.5200
	MAE	3.3519	2.7220
BA	MSE	39.3633	21.4275
	MAE	4.5106	3.4279

Finally, the LMS based forecasts might still contain additional explanatory power which could be a reward for the high computational cost incurred. To analyze this possibility, we apply the test proposed by Chong and Hendry (1986) to the forecasts obtained from LMS and OLS, respectively. Let $\hat{r}^e_{LMS,t}$ and $\hat{r}^e_{OLS,t}$ denote the conditional forecasts of the excess returns from the LMS and OLS estimates, and r^e_t the actual excess return in period t. Then estimation of the model

$$r^e_t = \gamma_0 + \gamma_1 \hat{r}^e_{LMS,t} + \gamma_2 \hat{r}^e_{OLS,t} + \nu_t \tag{5}$$

allows to test several hypotheses. If either γ_1 or γ_2 are equal to zero and $\gamma_0 = 0$, the forecast is unbiased. If $\gamma_1 = 0$, the forecast based on the OLS estimation dominates, i.e., the LMS based forecast does not provide additional information. For $\gamma_2 = 0$ we obtain dominance for the LMS based forecast. If both $\gamma_1 = 0$ and $\gamma_2 = 0$ has to be rejected, a combined forecast improves forecasting performance.

The results for the CAPM are again mixed. Only for the IBM stock, the null hypothesis that the forecasts based on the LMS estimates have no additional informational content can be rejected at the 5% level, while for the other stocks the OLS based forecast is found to dominate. However, one has to keep in mind that the test is linked to MSE, for which the OLS estimators should be more suitable. Thus, finding relevant additional information content (the parameter γ_1 has a value of close to 0.2 for IBM) demonstrates that considering alternative estimators might at least sometimes improve the predictive performance. The evidence changes when considering the multi factor model. There, for all stocks considered both γ_1 and γ_2 are significantly different from zero. Although the parameter values are typically smaller for γ_1, we find a clear evidence that the conditional forecast can be improved by combining the OLS based forecast with the LMS based forecast. Future research will focus on identifying the driving forces of this result and its robustness with regard to different sample length for the estimation periods.

5 Rate of Convergence

The results show that both optimization heuristics are able to solve the LMS estimation problem. We are also interested in comparing their rate of convergence. Typically, limits in time and computational resources make it unfeasible to obtain the global optimum in each run with certainty. However, by analyzing the distribution of outcomes for different parameter settings we can draw some conclusions on the convergence properties. Tables 3 and 4 provide a statistical summary of the results obtained by TA and DE for

various parameter settings. Again, we consider the first 200 observations for the IBM stock and the CAPM as our test case.

In our experiments, we calculate the LMS estimators by DE for nine combinations of population size $n_p = \{20, 50, 10, 200\}$ and number of generations $n_G = \{50, 100, 200, 1000\}$. The scaling factor F and the crossover rate CR are kept constant at 0.8 and 0.9, respectively. In the case of TA, the first parameter represents the number of rounds (corresponding to neighbourhood/threshold sequence reductions), n_R, and the second parameter represents the number of neighborhood search iterations n_S per round. Nine combinations of n_R and n_S are selected from $n_R = \{20, 50, 10, 200\}$ and $n_S = \{50, 100, 200, 1000\}$.

The implementation process of all heuristics are subject to random effects – the TA algorithm generates a random solution from a neighborhood, and the DE algorithm generates an initial population of random solutions. Furthermore, the selection of candidate solutions in each search step are random. In order to obtain some information on the effect of these stochastic components on the results, we repeat both algorithms 30 times for each set of parameter values and report the best value, the median, the worst value, the variance, the 5th percentile, the 90th percentile, and the frequency of the best value occuring in all 30 repetitions.

Looking first at the TA results, we can see that they improve significantly as the number of threshold sequence/neighbourhood reductions n_R and neighborhood iterations, n_S increase. However, the best results obtained by TA is not better than any of the DE results. Moreover, we observe that typically, the best results for a given parameter setting is found only once out of 30 restarts in each TA experiment. The results obtained by DE, which are shown in Table 4, contrast markedly with the results for TA. Here, we observe that DE converges close to the global optimum, and it achieves this in every restart when population size and generations are at least 50 and 100, respectively. Furthermore, it exhibits a very high convergence ratio in seven out of nine cases. The number of generations is the parameter which controls the consistency of the algorithm. Even with a population size of 20 the algorithm exhibits a high frequency of convergence when the number of generations is 100 or more. Using $n_p = 50$ and $n_G = 100$, a balance between variance and computational speed is attained.

The superior performance of DE is attributed to the fact that the search is run on a continuous search space. However, as pointed out above, the LMS estimation problem could also be interpreted as the search on a discrete search space. Then, the relative performance of TA might improve substantially as the results by Fitzenberger and Winker (2007) for the related problem of censored quantile regression show.

Table 3: Descriptive statistics for 30 repetitions of the LMS estimation using TA.

n_R	n_S	**Best value**	**Median**	**Worst value**	**Var**	**q5%**	**q90%**	**Freq**
20	50	4.9961-005	5.0563-005	5.1557-005	1.7399-013	4.9961-005	5.1213-005	1
20	100	4.9959-005	5.0165-005	5.0913-005	5.8351-014	4.9959-005	5.0532-005	1
20	200	4.9959-005	5.0095-005	5.0745-005	3.0213-014	4.9959-005	5.0327-005	1
20	1000	4.9936-005	4.9971-005	5.0113-005	1.9420-015	4.9936-005	5.0052-005	1
50	50	4.9975-005	5.0653-005	5.1467-005	1.8137-013	4.9975-005	5.1151-005	1
50	100	4.9945-005	5.0106-005	5.0735-005	3.1858-014	4.9945-005	5.0403-005	1
100	50	4.9950-005	5.01441-005	5.0552-005	2.3537-014	4.9950-005	5.0324-005	1
100	100	4.9945-005	5.0016-005	5.0309-005	9.6951-015	4.9945-005	5.0174-005	1
200	200	4.9939-005	4.9961-005	5.0041-005	8.8922-016	4.9939-005	5.0027-005	1

Table 4: Descriptive statistics for 30 repetitions of the LMS estimation using DE.

n_p	n_G	**Best value**	**Median**	**Worst value**	**Var**	**q5%**	**q90%**	**Freq**
20	50	4.9935-005	4.9936-005	5.0934-005	6.3944-014	4.9935-005	5.0025-005	15
20	100	4.9935-005	4.9935-005	5.0934-005	9.2912-014	4.9935-005	5.0434-005	27
20	200	4.9935-005	4.9935-005	5.0934-005	3.3265-014	4.9935-005	4.9935-005	29
20	1000	4.9935-005	4.9935-005	5.0934-005	6.4223-014	4.9935-005	4.9936-005	27
50	50	4.9935-005	4.9935-005	4.9974-005	6.0958-017	4.9935-005	4.9941-005	18
50	100	4.9935-005	4.9935-005	4.9935-005	3.7887-028	4.9935-005	4.9935-005	30
100	50	4.9935-005	4.9935-005	4.9936-005	2.367e-019	4.9935-005	4.9936-005	25
100	100	4.9935-005	4.9935-005	4.9935-005	3.0485-030	4.9935-005	4.9935-005	30
200	200	4.9935-005	4.9935-005	4.9935-005	4.7501-041	4.9935-005	4.9935-005	30

Nevertheless, it is useful to report how much more efficient DE is than TA for the given problem formulation. DE and TA are both different search methods, and the main computational burden for search heuristics is calculating the objective function. Therefore, we use this as the measure for efficiency. In each TA paired DE experiment, i.e., in the corresponding lines of Tables 3 and 4, we calculated the objective function the same number of times. Considering, e.g., for TA the experiment with $n_R = 20$ and $n_S = 1000$ (line 4), we find that it has a lower variance than the paired application of DE. Nevertheless, the best result for TA is still slightly worse than the best results obtained by DE. Again, this difference in efficiency might be attributed to the difficulties TA faces in the fine tuning of continuous variables.

To summarize, the results we obtain indicate the superiority of DE in terms of consistency and efficiency for LMS estimation.

6 Conclusion and Further Work

The LMS estimator is considered for obtaining robust estimators of the parameters of the CAPM and a multi factor model. It is shown that optimization heuristics like TA and DE are suitable to solve the resulting optimization problem. Thereby, DE appears to be more efficient when the underlying discrete structure of the search space is ignored.

In fact, efficient implementations of DE allow a fast and reliable estimation of the parameters of both models by LMS. This is demonstrated by a rolling window analysis on a sample of six publicly traded firms with daily data for the CAPM (1970 - 2006) and monthly data for the multi factor model (1970 - 2008). The results indicate that the estimates obtained by LMS differ substantially from those resulting from OLS. However, the LMS estimates do not exhibit less variation as might have been expected from the outlier related argument. Furthermore, the relative performance of both estimators in a simple one-period-ahead conditional forecasting experiment is mixed. In most cases, both the MSE and the MAE are smaller for the conditional forecasts based on the OLS estimation. However, it is shown that the forecasts based on the LMS estimators provide additional informational content. Thus, it might be justified to incure the additional computational load for obtaining the LMS estimators.

Some extensions of the paper are straightforward based on the results presented. First, the method should be applied to different data sets, e.g., stock returns from other stock indices or stock markets. Furthermore, it would be of interest to identify in more detail the situations when the estimation and forecast based on LMS outperforms OLS and vice versa.

References

Barreto, H. and D. Maharry (2006). Least median of squares and regression through the origin. *Computational Statistics & Data Analysis* **50**(6), 1391–1397.

Chan, L.K.C. and J. Lakonishok (1992). Robust measurement of beta risk. *Journal of Financial and Quantitative Analysis* **27**, 265–282.

Chong, Y.Y. and D.F. Hendry (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies* **53**(175), 671–690.

Cornell, B. and J.K. Dietrich (1978). Mean-absolute-deviation versus least-squares regression estimation of beta coefficients. *Journal of Financial and Quantitative Analysis* **13**, 123–131.

Dueck, G. and P. Winker (1992). New concepts and algorithms for portfolio choice. *Applied Stochastic Models and Data Analysis* **8**, 159–178.

Dueck, G. and T. Scheuer (1990). Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. *J. Comp. Phys.* **90**, 161–175.

Fama, E.F. and K.R. French (1992). The cross-section of expected stock returns. *Journal of Finance* **47**, 427–465.

Fama, E.F. and K.R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**, 3–56.

Fitzenberger, B. and P. Winker (2007). Improving the computation of censored quantile regressions. *Computational Statistics & Data Analysis* **52**(1), 88–108.

Gilli, M. and P. Winker (2007). Editorial - 2nd special issue on applications of optimization heuristics to estimation and modelling problems. *Computational Statistics & Data Analysis* **52**(1), 2–3.

Gilli, M. and P. Winker (2008). A review of heuristic optimization methods in econometrics. Research Paper 08-12. Swiss Finance Institute. Geneva.

Gilli, M., D. Maringer and P. Winker (2008). Applications of heuristics in finance. In: *Handbook of Information Technoloy in Finance* (D. Seese, C. Weinhardt and F. Schlottmann, Eds.). Chap. 26, pp. 635–653. International Handooks on Information Systems. Springer. Berlin.

Knez, P.J. and M.J. Ready (1997). On the robustness of size and book-to-market in cros–sectional regressions. *The Journal of Finance* **52**(4), 1355–1382.

Krink, T., S. Paterlini and A. Restic (2007). Using differential evolution to improve the accuracy of bank rating systems. *Computational Statistics & Data Analysis* **52**(1), 68–87.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics & Statistics* **47**(1), 13–37.

Maringer, D. (2005). Distribution assumptions and risk constraints in portfolio optimization. *Computational Management Science* **2**(2), 139–153.

Maringer, D. and M. Meyer (2008). Smooth transition autoregressive models - new approaches to the model selection problem. *Studies in Nonlinear Dynamics & Econometrics* **12**(1), 5.

Markowitz, H.M. (1952). Portfolio selection. *Journal of Finance* **7**(1), 77–91.

Martin, R.D. and T. Simin (2003). Outlier resistant estimates of beta. *Financial Analysts Journal* **59**, 56–69.

Mayo, M.S. and J.B. Gray (1997). Elemental subsets: The building blocks of regression. *The American Statistician* **51**(2), 122–129.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* **34**(4), 768–783.

Price, K. V., R. M. Storn and J. A. Lampinen (2005). *Differential Evolution: A Practical Approach to Global Optimization*. Springer, Germany.

Ronchetti, E. and M. Genton (2008). Robust prediction of beta. In: *Computational Methods in Financial Engineering* (E. Kontoghiorghes, B. Rustem and P. Winker, Eds.). pp. 147–161. Springer. Berlin.

Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871–880.

Rousseeuw, P.J. and A.M. Leroy (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

Rousseeuw, P.J. and J. Wagner (1994). Robust regression with a distributed intercept using least median of squares. *Computational Statistics & Data Analysis* **17**, 65–76.

Sharpe, W. (1971). Mean-absolute deviation characteristic lines for securities and portfolios. *Management Science* **18**, 1–13.

Sharpe, W.F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* **19**(3), 425–442.

Simin, T. (2008). The poor predictive performance of asset pricing models. *Journal of Financial and Quantitative Analysis* **43**(2), 355–380.

Specht, K. and P. Winker (2008). Portfolio optimization under VaR constraints based on dynamic estimates of the variance-covariance matrix. In: *Computational Methods in Financial Engineering* (E. Kontoghiorghes, B. Rustem and P. Winker, Eds.). pp. 73–94. Springer. Berlin.

Storn, R. and K. Price (1997). Differential evolution: a simple and efficient adaptive scheme for global optimization over continuous spaces. *J. Global Optimization* **11**, 341–359.

Winker, P. (2001). *Optimization Heuristics in Econometrics: Applications of Threshold Accepting*. Wiley, New York.

Winker, P. and D. Maringer (2007*a*). The hidden risks of optimizing bond portfolios under VaR. *Journal of Risk* **9**(4), 1–19.

Winker, P. and D. Maringer (2007*b*). The threshold accepting optimisation algorithm in economics and statistics. In: *Optimisation, Econometric and Financial Analysis* (Erricos J. Kontoghiorghes and C. Gatu, Eds.). Vol. 9 of *Advances in Computational Management Science*. pp. 107–125. Springer.

Winker, P. and K.-T. Fang (1997). Application of threshold accepting to the evaluation of the discrepancy of a set of points. *SIAM Journal on Numerical Analysis* **34**(5), 2028–2042.

Yang, Z., Z. Tian and Z. Yuan (2007). GSA-based maximum likelihood estimation for threshold vector error correction model. *Computational Statistics & Data Analysis* **52**(1), 109–120.

Zaman, A., P.J. Rousseeuw and M. Orhan (2001). Econometric applications of high-breakdown robust regression techniques. *Economics Letters* **71**, 1–8.