Winker, Peter; Maringer, Dietmar

Working Paper

The convergence of optimization based estimators : theory and application to a GARCH-model

Discussion Paper, No. 2005,004E

Provided in Cooperation with:
University of Erfurt, Faculty of Economics, Law and Social Sciences

Suggested Citation: Winker, Peter; Maringer, Dietmar (2005) : The convergence of optimization based estimators : theory and application to a GARCH-model, Discussion Paper, No. 2005,004E, Universität Erfurt, Staatswissenschaftliche Fakultät, Erfurt

This Version is available at:
https://hdl.handle.net/10419/23941

Staatswissenschaftliche Fakultät

Faculty of Economics,
Law and Social Sciences

Discussion Paper No.: 2005-004E

The Convergence of Optimization Based Estimators: Theory and Application to a GARCH-Model

Peter Winker, Dietmar Maringer

ISSN 1610-9198 (Print)

ISSN 1610-918X (Internet)

The Convergence of Optimization Based Estimators: Theory and Application to a GARCH-Model*

Peter Winker† Dietmar Maringer‡

August 22, 2005

Abstract

The convergence of estimators, e.g. maximum likelihood estimators, for increasing sample size is well understood in many cases. However, even when the rate of convergence of the estimator is known, practical application is hampered by the fact, that the estimator cannot always be obtained at tenable computational cost.

This paper combines the analysis of convergence of the estimator itself with the analysis of the convergence of stochastic optimization algorithms, e.g. threshold accepting, to the theoretical estimator. We discuss the joint convergence of estimator and algorithm in a formal framework.

An application to a GARCH-model demonstrates the approach in practice by estimating actual rates of convergence through a large scale simulation study. Despite of the additional stochastic component introduced by the use of an optimization heuristic, the overall quality of the estimates turns out to be superior compared to conventional approaches.

Keywords: GARCH; Threshold Accepting; Optimization Heuristics; Convergence.
JEL classification: C22, C63.

*We are indebted to Manfred Gilli for valuable comments on a preliminary draft of this paper.
†Department of Economics, Law and Social Sciences, University of Erfurt
‡Department of Economics, Law and Social Sciences, University of Erfurt

1 Introduction

The convergence of estimators, e.g. maximum likelihood estimators, for increasing sample size is well understood in many cases. However, even when the rate of convergence of the estimator is known, practical application is hampered by the fact, that the estimator cannot always be obtained at tenable computational cost. In fact, the literature mentions many estimation problems, where standard optimization methods fail to provide a reliable approximation to the theoretical estimator. Examples include switching regression models (Dorsey and Mayer, 1995; Clements and Krolzig, 1998), censored quantile regression (Fitzenberger, 1997; Fitzenberger and Winker, 1998) or the GARCH-model (Brooks *et al.*, 2001). Even for simpler problems, standard software might fail to provide adequate results (McCullough and Vinod, 1999; McCullough and Wilson, 1999; McCullough and Wilson, 2002; McCullough and Wilson, 2005). Often, this failure of standard methods is not due to a suboptimal implementation of the algorithms, but results from the inherent computational complexity of the problems and has to be taken as given (Winker, 2001, pp. 57ff).

However, if the theoretical estimator has to be replaced by some numerical approximation, the actual rate of convergence might differ from the theoretical one. In fact, if the implementation of the estimator is such that it will not converge to the theoretical estimator with the sample size growing to infinity, the convergence properties of the estimator get lost. Unfortunately, many real life implementations of complex estimators cannot guarantee to result in the true theoretical estimator or, at least, a close approximation. Furthermore, typically, the algorithms are not constructed in a way to offer some options for a satisfying convergence as they are built with the purpose to obtain the theoretical estimator. Thus, if these methods work fine, the theoretical convergence results apply, if they fail, no statement on convergence can be provided.

The picture changes when the algorithm for calculating the estimator itself might be subject to a stochastic analysis. This is the case, e.g. for optimization heuristics like genetic algorithms or local search heuristics. In particular, when it can be proven that the result found by the heuristic converges to the theoretical estimator with an increasing number of iterations, a joint convergence analysis becomes feasible. In this contribution, we consider the threshold accepting heuristic, for which Althöfer and Koschnik (1991) provide such a convergence result. A first detailed analysis of the stochastic properties of this algorithm in an application to experimental design is provided by Winker (2005). Here, we consider a standard estimation problem, namely the maximum likelihood estimation of the parameters

of a GARCH-model. The estimation problem and the application of threshold accepting to this problem is described by Maringer (2005, pp. 63ff). The aim of this contribution is to derive and analyze the joint convergence properties of the optimization algorithm and the maximum likelihood estimator. In particular, by means of a large scale simulation study, we estimate the number of iterations of the optimization algorithm as a function of the sample size required to obtain a standard rate of convergence for the actual parameter estimates.

The paper is structured as follows. Section 2 provides a formal framework for the analysis of joint convergence of the calculation of the estimator and the estimator itself. The framework is applied to a GARCH-model in Section 3, where a large scale simulation study is introduced. The results of this simulation study and its implication on the actual convergence properties of the GARCH-estimator are provided in Section 4. Section 5 summarizes the main findings and provides an outline for further research.

2 Convergence of Optimization Based Estimators

2.1 Notation

Before turning to the discussion of the convergence properties of optimization based estimators, we have to introduce some notation. We assume that the true model for the data generating process is known except for the values of a number of parameters collected in the true parameter vector $\boldsymbol{\psi}^{\mathrm{TR}}$. In particular, we will not consider issues related to model misspecification. For a given data sample consisting of T observations, let $\boldsymbol{\psi}^{\mathrm{ML},T}$ denote the value of the theoretical estimator, e.g. the maximum likelihood estimator for the GARCH-model. This vector cannot be observed unless a deterministic algorithm is available which provides the estimator with certainty. Such a situation is given, e.g., for the ordinary least squares estimator when the problem size is not too large and the scaling and multicollinearity of the explanatory variables is not too extreme. However, as pointed out before, this condition is not fulfilled for the GARCH-model when relying on standard optimization tools (Brooks *et al.*, 2001).

In contrast, when a stochastic optimization heuristic like threshold accepting has to be used to obtain an approximation of the estimator, only one or several realizations of this stochastic procedure can be observed. The quality of these realizations will depend on the computational effort spent on the optimization process. For threshold accepting, the number of local search steps or iterations I

provides an adequate measure of this computational effort. Thus, if the optimization is run R times with I iterations for each run, we obtain R approximations of $\boldsymbol{\psi}^{\mathrm{ML},T}$, which are denoted as $\boldsymbol{\psi}^{T,I,r}$, where $r = 1,\ldots,R$.

Now, the two aspects of convergence of optimization based estimators can be discussed. First, asymptotic consistency of the theoretical estimator $\boldsymbol{\psi}^{\mathrm{ML},T}$ with regard to sample size T has to be established. This is the usual task considered in econometric analysis. Second, we have to demonstrate that based on the approximations $\boldsymbol{\psi}^{T,I,r}$ it is possible to obtain convergence in probability towards $\boldsymbol{\psi}^{\mathrm{ML},T}$ as I goes to infinity. Finally, in a third step we have to show that both results can be combined to obtain a convergence result for the estimator found by the threshold accepting implementation. In particular, we have to provide a relationship $I(T)$ resulting in a convergence in probability of an estimate based on the $\boldsymbol{\psi}^{T,I,r}$ towards the true parameter vector $\boldsymbol{\psi}^{\mathrm{TR}}$.

This task is slightly complicated by the fact that in a stochastic optimization approach, it is not optimal to spend all available computational resources on a single run, i.e. $R = 1$, but rather to allow for a small number of runs R and to use the best result out of these runs corresponding to the first order statistic of the objective function, e.g. the likelihood function (Winker, 2005).

2.2 Convergence of the Estimator

The maximum likelihood estimator of the GARCH-model $\boldsymbol{\psi}^{\mathrm{ML},T}$ converges with the standard rate $\sqrt{T}$ to the true parameter vector $\boldsymbol{\psi}^{\mathrm{TR}}$ and is asymptotically normally distributed if the usual regularity conditions are satisfied (Herwartz, 2004, p. 202). An equivalent expression of this convergence result is the following: For any given $\delta > 0$ and $\varepsilon > 0$, there exists a sample size $T(\delta,\varepsilon)$ such that for any $T \geq T(\delta,\varepsilon)$ we find

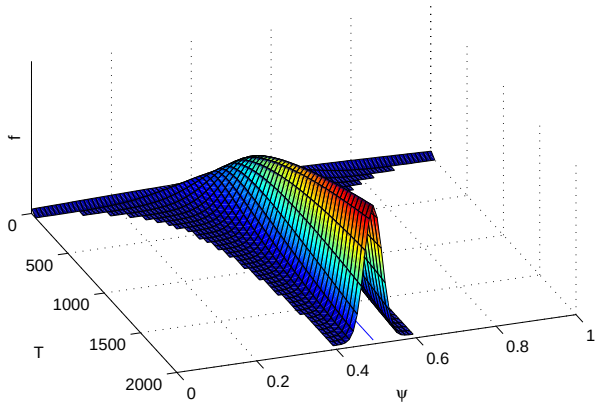
$$\mathrm{P}(|\boldsymbol{\psi}^{\mathrm{ML},T} - \boldsymbol{\psi}^{\mathrm{TR}}| < \varepsilon) > 1 - \delta\,. \tag{1}$$

In fact, for given δ, asymptotically, T has to be chosen proportional to $1/\varepsilon^2$ to obtain (1). Figure 1 provides a stylized illustration of this convergence property of the estimator with regard to increasing T.

2.3 Convergence of Threshold Accepting

Suggested by Dueck and Scheuer (1990), threshold accepting is a heuristic optimization method where a solution is repeatedly modified and updated in a stochas-

Figure 1: Convergence of the ML estimator with regard to T



tic fashion.[1] Consequently, repeated runs of the optimization heuristic on a single problem instance will result in a distribution of results $\boldsymbol{\psi}^{T,I,r}$.

According to the convergence result for threshold accepting obtained by Althöfer and Koschnik (1991), there exist suitable parameters for the threshold accepting implementation such that the global optimum of the objective function can be approximated at arbitrary accuracy with any fixed probability close to one by increasing the number of iterations. If the search of parameters $\boldsymbol{\psi}^{T,I,r}$ is restricted to a compact set, the continuity of the likelihood function allows for the following conclusion from this convergence result: For any given $\delta > 0$ and $\varepsilon > 0$, there exists a number of iterations $I(\delta,\varepsilon)$ such that

$$\mathrm{P}(|\boldsymbol{\psi}^{T,I,r} - \boldsymbol{\psi}^{\mathrm{ML},T}| < \varepsilon) > 1 - \delta \tag{2}$$

for any $r = 1,\ldots,R$. Obviously, the convergence of the first order statistic of $\boldsymbol{\psi}^{T,I,r}$, $r = 1,\ldots,R$ will also satisfy this condition – potentially for a smaller value of I. Unfortunately, the theoretical convergence result does not allow to derive a general result on the required number of iterations $I(\delta,\varepsilon)$. Consequently, it will be left to the analysis of our empirical implementation to demonstrate that $I(\delta,\varepsilon)$ can be chosen to be a function of $T(\delta,\varepsilon)$ growing at a less than linear rate.

[1] For a more detailed presentation of this method and applications in economics and statistics, see Winker (2001).

2.4 Joint Convergence

The stochastic feature of the optimization heuristic might appear like a drawback on first sight as compared to standard optimization tools. However, a combination of the convergence results for estimator and optimization allows to derive a joint convergence result, which, in general, cannot be obtained for deterministic procedures.[2]

Let $\varepsilon > 0$ be a predefined required level of accuracy of the estimator with regard to the true parameter value. Furthermore, let $\delta > 0$ denote an admissible (though small) probability for missing this level of accuracy. Then, according to (1), we find $T(\delta/2, \varepsilon/2)$ such that

$$\mathrm{P}(|\boldsymbol{\psi}^{\mathrm{ML},T} - \boldsymbol{\psi}^{\mathrm{TR}}| < \varepsilon/2) > 1 - \delta/2\,. \tag{3}$$

Furthermore, using (2) for an adequate number of iterations $I(T(\delta/2, \varepsilon/2))$, we find

$$\mathrm{P}(|\boldsymbol{\psi}^{T,I,R} - \boldsymbol{\psi}^{\mathrm{ML},T}| < \varepsilon/2) > 1 - \delta/2\,, \tag{4}$$

where $\boldsymbol{\psi}^{T,I,R}$ denotes the estimate corresponding to the best result out of R replications of the threshold accepting heuristic. Remember that the empirical application in section 3 will demonstrate that in practice $I(T(\delta/2, \varepsilon/2))$ can be bounded by a linear function of $T(\delta/2, \varepsilon/2)$ rendering a real implementation feasible.

Combining (3) and (4), we find

$$\mathrm{P}(|\boldsymbol{\psi}^{T,I,R} - \boldsymbol{\psi}^{\mathrm{TR}}| < \varepsilon) > 1 - \delta\,, \tag{5}$$

i.e. convergence of the heuristic optimization based estimator to the true parameter value for T going to infinity and I going to infinity as a function of T.

3 Application to GARCH-Model

3.1 Model and Data for the Computational Study

As a benchmark implementation for assessing the performance of the estimation method in practice, we consider the basic GARCH(1,1) model

$$r_t = \psi_0 + e_t \;\text{ with }\; e_t \sim N(0, \sigma_t^2)\,, \tag{6}$$

[2] The obvious exceptions are those estimates which can be obtained with certainty by means of a deterministic algorithm given the available computational resources.

where

$$\sigma_t^2 = \psi_1 + \psi_2 e_{t-1}^2 + \psi_3 \sigma_{t-1}^2 \,. \tag{7}$$

$$\tag{8}$$

For the empirical application, we refer to the estimates obtained by Bollerslev and Ghysels (1996) based on 1974 daily observations for the changes in the German mark / British pound exchange rate. Their maximum likelihood estimates of the parameters of the GARCH(1,1) are the following (using our notation):

$$\boldsymbol{\psi}^{\mathrm{TR}} = [\psi_0^{\mathrm{TR}} \ldots \psi_3^{\mathrm{TR}}] = [\ -0.00619041 \quad 0.0107613 \quad 0.153134 \quad 0.805974\ ]. \tag{9}$$

We use this model and parameters for a data generating process and produce 100 time series each consisting of 2100 observations.[3] For the computational study, we then removed the first 100 observations (which were used as a forerun to allow the process to swing in). From the remaining series, we analyzed the first T observations with T = 50, 100, 200, 400, 1000, and 2000.

3.2 The Optimization Heuristic

For finding the parameters that maximize the loglikelihood function of the GARCH model,

$$L = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left(\ln(\sigma_t^2) + \frac{e_t^2}{\sigma_t^2}\right), \tag{10}$$

we use the threshold accepting implementation suggested in Maringer (2005, chapter 2). First, a random initial solution for $\boldsymbol{\psi}$ is generated. The only constraint on this initial solution is that the values of all of parameters must be within certain limits: In order to avoid negative values for σ_t^2, the parameters ψ_1, ψ_2, and ψ_3 must not be negative. Also the ARCH and GARCH parameters ψ_2 and ψ_3, respectively, must not exceed 1.[4] For ψ_0, values within the interval $[-1;+1]$ were accepted which appeared to be a sufficiently generous range for daily changes in the exchange rates.

In each of the following iterations, a new solution $\boldsymbol{\psi}^*$ is generated. This is done by changing one element of the vector $\boldsymbol{\psi}$ by adding a random term δ to its

[3] This approach is based on the idea of data based Monte Carlo simulation as introduced by Ho and Sørensen (1996).

[4] A more rigorous constraint would have been $\psi_2 + \psi_3 \leq 1$.

current value while keeping the other elements unchanged; which of the elements is changed is also determined randomly. The error term δ is uniformly distributed within a range $[-u_i; u_i]$. u_i therefore defines a neighborhood around the current value of the parameter. As indicated by the index i, u_i is subject to change over the iterations. It proved favorable when the neighborhood is narrowed down during the optimization process by decreasing u_i in regular intervals; Table 1 summarizes the values which appeared most efficient in preceding experiments for this problem and were therefore chosen for our implementation. If the changed value would exceed one of the lower or upper bounds used for its initialization, the value is set equal to this bound.

number of iterations, I	1 000	5 000	10 000	25 000	50 000	100 000
initial value for u	.05	.025	.025	.025	.025	.01
terminal value for u	.005	.0025	.00125	.00125	.00125	.001

Table 1: Parameters for the Threshold Accepting implementation

After generating the new solution $\boldsymbol{\psi}^*$ by modifying $\boldsymbol{\psi}$, the value of the log-likelihood function for this new parameter combination, L^*, is determined and compared to the one of the previous $\boldsymbol{\psi}$, L. If $L^* > L$, the random change caused an improvement, and the modified parameter set $\boldsymbol{\psi}^*$ replaces the previous solution $\boldsymbol{\psi}$. Contrariwise, $L^* < L$ means that the random change in $\boldsymbol{\psi}$ degraded the solution. In order to overcome local optima, however, the change is kept anyway if the impairment is not too severe, i.e., if it does not exceed a given threshold τ_i. In short, the random change will therefore be kept if $L^* + \tau_i > L$, otherwise it will be undone and $\boldsymbol{\psi}$ is kept as the current solution. The initial value for this threshold, τ_0, is set to 0.01.[5] During the optimization process, it is linearly lowered in regular intervals towards zero. Hence, the algorithm is rather tolerant in accepting impairments in the beginning, yet strict during the last iterations.

This local neighborhood search was repeated over a given number of iterations, I. The algorithm then reports the parameter vector $\boldsymbol{\psi}$ corresponding to the highest value of the likelihood function L found in any of these iterations. The algorithm is implemented using the Delphi (Version 7) programming environment and executed on Pentium IV machines, and the CPU time per optimization run is in the range of less than a second (for short data series and a low number of

[5] In preliminary experiments, alternative values depending on I were tested. It turned out, however, that there is hardly any additional gain by this flexibility when u is variable.

iterations) up to approximately 20 seconds (for process with 2000 observations and using 100000 iterations).

4 Results

4.1 Notation

As we consider 6 different values for both T and I and 100 data series, this adds up to 3 600 different optimization problems. For each of these problems, the algorithm was run approximately $R \approx 1700$ times, resulting in a total of 6 146 393 reported solutions. For evaluation purposes, we then computed the mean squared deviation between the reported parameters and the true (TR) and the maximum likelihood parameters (ML), respectively:

$$MSD_p^{\mathrm{TR},d,T,I} = \frac{1}{R} \cdot \sum_{r=1}^{R} \left(\psi_p^{d,T,I,r} - \psi_p^{\mathrm{TR}} \right)^2 \tag{11}$$

$$MSD_p^{\mathrm{ML},d,T,I} = \frac{1}{R} \cdot \sum_{r=1}^{R} \left(\psi_p^{d,T,I,r} - \psi_p^{\mathrm{ML},d,T} \right)^2 \tag{12}$$

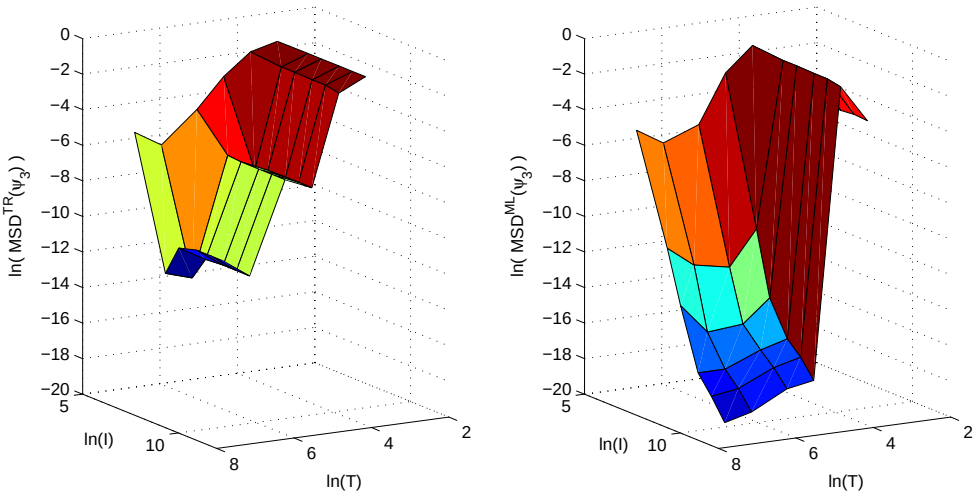
where $\psi_p^{d,T,I,r}$ is the p-th element of the optimal parameter vector for the data series d with T observations reported in the r-th run and found within I iterations. While the true parameters $\boldsymbol{\psi}^{\mathrm{TR}}$ are known from the data generating process and are the same for all processes d and lengths T (see equation (9)), the (supposed) maximum likelihood parameters $\boldsymbol{\psi}^{\mathrm{ML},d,T}$ are the best results for process d and length T reported in any of the runs or by the Matlab toolbox (as described in section 4.3). Figure 2 illustrates the results for the *MSD*'s for the GARCH parameter ψ_3 for one specimen data series d as a function of the sample size T and the number of iterations I used in the threshold accepting implementation.

4.2 Convergence Behavior

4.2.1 Convergence of the Estimator

In section 2.2 it was stated that the maximum likelihood parameters will converge to the true values when the length of the data series, T, is increased. Figure 3 depicts the median and the 25% and 75% quantiles, respectively, of the optimal parameters for the 100 data series in dependence of T. Akin to the stylized graph

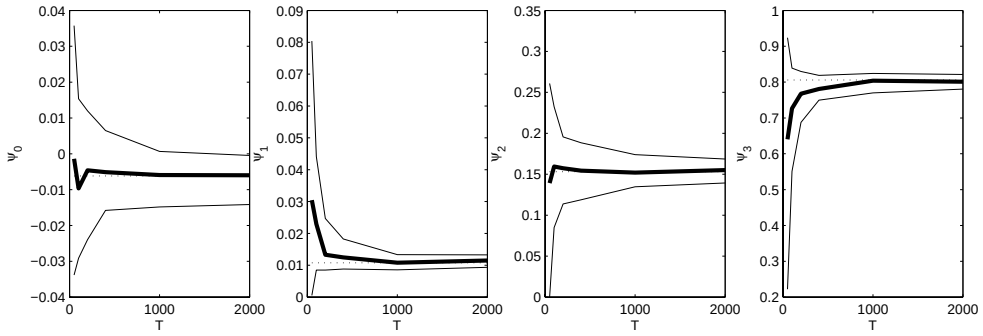
Figure 2: MSD^{TR} (left) and MSD^{ML} (right) for ψ_3 for one specimen data series



in Figure 1, this illustrates the relationship between the number of observations, T, and the range within which the maximum likelihood estimators are likely to be found. In particular for the parameters ψ_1 and ψ_3, it can be noted that they are not symmetrically distributed when T is small. This can be partially attributed to the imposed limits on these parameters ($0 \leq \psi_i \leq 1$ for $i = 1,..,3$). In any case, however, an increase in T lets the medians eventually converge to the true values, the range between the upper and lower quantiles narrows down – and, thus, the mean squared deviations from the true values decrease.

This convergence can be expected to show up also when individual data series are considered. The left graph in Figure 2 suggests a linear relationship between the logs of T and MSD^{TR}, i.e., that the maximum likelihood parameters tend to converge to the true values when a time series is prolonged. The slope of the linear relationship provides an estimate of the rate of convergence. However, this convergence is not necessarily smooth: If the added observations contain an unusually high number of outliers, an increase in T might well drive the maximum likelihood estimators away from their true values (i.e., increases the MSD^{TR}). This effect will show primarily when the number of observations is still rather low; nonetheless it might also appear in longer time series, but will eventually disappear as T goes to infinity.

Figure 3: Median (thick line) and 25% and 75% quantiles (thin lines) of the maximum likelihood estimators of the 100 data series



To test for this convergence property, we estimate the linear relationship between the logs of the mean squared deviations from the true values and the number of observations. In order to isolate effects of the optimization heuristic, we group the data by the number of the algorithm's iterations, I, and then estimate the parameters of the model

$$\ln\left(MSD_p^{\mathrm{TR},d,T,I}\right) = a_p^{d,I} + b_p^{d,I} \cdot \ln(T). \tag{13}$$

for each data series $d = 1 \ldots 100$.

Table 2 summarizes the aggregate results and some statistics. As has been stated in section 2.3 (and will be confirmed by our empirical results in section 4.2.2), the optimization algorithm will produce more reliable results when it is allowed a higher number of iterations. Hence, the main focus should be on the results for $I = 100000$; however, it is safe to say that if the algorithm is conceded at least 5000 iterations, the conclusions are virtually the same.[6]

On average, the mean rate of convergence of $MSD_p^{\mathrm{TR},d,T,I}$ as a function of T is found to be approximately of the order of $\frac{1}{T}$,[7] and it is even faster for ψ_1 and ψ_3. The relationship is also supported by the high average R^2's. However, the relatively large standard deviations of the parameters $b_p^{d,I}$ indicate that their values (and thus the convergence rates) can differ substantially between different realizations of the data generating process. We found two main reasons for this. As

[6]The joint effects of the number of iterations and the number of observations will be discussed in section 4.2.3.

[7]This corresponds to the usual rate of convergence of $\frac{1}{\sqrt{T}}$ for the parameters.

I	Values for b_p^I, averaged over d			
	$MSD^{\mathrm{TR},I}(\psi_0)$	$MSD^{\mathrm{TR},I}(\psi_1)$	$MSD^{\mathrm{TR},I}(\psi_2)$	$MSD^{\mathrm{TR},I}(\psi_3)$
1000	-0.921	-1.171	-0.394	-0.936
5000	-0.969	-1.950	-1.143	-1.725
10000	-0.969	-1.985	-1.159	-1.716
25000	-0.967	-1.949	-1.157	-1.690
50000	-0.965	-1.912	-1.160	-1.669
100000	-0.962	-1.890	-1.159	-1.643

I	Standard deviation of the reported values $b_p^{d,I}$			
1000	0.647	0.641	0.404	0.409
5000	0.724	0.824	0.664	0.663
10000	0.728	0.804	0.698	0.742
25000	0.730	0.815	0.723	0.788
50000	0.734	0.821	0.726	0.796
100000	0.737	0.830	0.731	0.789

I	Fraction of values for $b_p^{d,I}$ significantly different from 0 (5%)			
1000	0.31	0.50	0.26	0.63
5000	0.26	0.60	0.43	0.62
10000	0.26	0.58	0.44	0.58
25000	0.26	0.53	0.42	0.57
50000	0.26	0.51	0.42	0.58
100000	0.26	0.53	0.42	0.58

I	Corresponding values for R^2, averaged over d			
1000	0.429	0.599	0.397	0.677
5000	0.429	0.646	0.583	0.671
10000	0.428	0.640	0.576	0.652
25000	0.429	0.624	0.567	0.639
50000	0.429	0.617	0.569	0.639
100000	0.428	0.615	0.567	0.639

Table 2: The influence of the number of observations, T, when the maximum number of iterations, I, is fixed

mentioned above, when the data series is extended and the additional observations contain outliers, this might increase the gap between $\boldsymbol{\psi}^{\mathrm{ML}}$ and $\boldsymbol{\psi}^{\mathrm{TR}}$; a further increase in the number of observations, however, will eventually outbalance this effect. However, the convergence rate is also low when there are hardly any outliers within the whole of the data series; in this case, $\boldsymbol{\psi}^{\mathrm{ML}}$ is already close to $\boldsymbol{\psi}^{\mathrm{TR}}$ for small T – and will remain small when T is increased. In these cases, the influence of T on MSD^{TR} will vanish and $b_p^{d,I}$ will not be significantly different from 0.

On the other hand, if the extreme values of a data series are concentrated in the first observations, then for short T, $\boldsymbol{\psi}^{\mathrm{ML}}$ will differ substantially from $\boldsymbol{\psi}^{\mathrm{TR}}$. Adding further observations are then likely to quickly drive the optimal parameters $\boldsymbol{\psi}^{\mathrm{ML}}$ towards $\boldsymbol{\psi}^{\mathrm{TR}}$, and the convergence rate will be substantially above average. Thus, according to our interpretation of the simulation results, the high standard deviation is mainly due to small sample effects.

4.2.2 Convergence of Threshold Acceptance

When analyzing the convergence of $\boldsymbol{\psi}^{\mathrm{ML}}$, a crucial question is how reliable the optimal (i.e., maximum likelihood) parameters are identified in the first place. When using threshold accepting for the optimization, the central influencing factor on the reliability is the number of iterations per run: the more iterations, the more time the algorithm is conceded to find the optimum. Under the assumption of a linear relationship between the logs of the MSD^{ML} and the number of iterations, I, a model of the type

$$\ln\left(MSD_p^{\mathrm{ML},d,T,I}\right) = a_p^{d,T} + b_p^{d,T} \cdot \ln(I) \tag{14}$$

can be estimated for each data series d and fixed length T. Table 3 summarizes the mean values for b_p^T and some statistics. The results confirm the previous considerations: in particular for long data series, the number of iterations has a negative influence on MSD^{ML}. Overall, the mean rate of convergence of $MSD^{\mathrm{ML},T}$ as a function of I is found to be of the order $\frac{1}{I}$ or faster. Though the convergence rate differs between data series, these differences diminish the longer the time series become, as can be seen from the standard deviations for the reported values of $b_p^{d,T}$. Also, in many (or, for large T, virtually all) cases, this relationship is statistically significant. The high average R^2's indicate that I is the main contributor to the deviation between the reported and the Maximum likelihood parameters for the GARCH model.

T	Values for b_p^T, averaged over d			
	$MSD^{\mathrm{ML},T}(\psi_0)$	$MSD^{\mathrm{ML},T}(\psi_1)$	$MSD^{\mathrm{ML},T}(\psi_2)$	$MSD^{\mathrm{ML},T}(\psi_3)$
50	-0.958	-1.196	-1.328	-1.094
100	-1.257	-1.829	-1.669	-1.734
200	-1.527	-2.561	-2.207	-2.520
400	-1.510	-2.841	-2.518	-2.794
1000	-1.514	-3.026	-2.826	-3.005
2000	-1.366	-3.107	-2.948	-3.099

T	Standard deviation of the reported values $b_p^{d,T}$			
50	0.896	1.140	1.005	1.044
100	0.809	1.265	1.001	1.173
200	0.533	0.962	0.757	0.916
400	0.389	0.762	0.570	0.712
1000	0.300	0.362	0.245	0.324
2000	0.250	0.228	0.166	0.202

T	Fraction of values for $b_p^{d,T}$ significantly different from 0 (5%)			
50	0.65	0.66	0.52	0.69
100	0.79	0.84	0.73	0.80
200	0.97	0.95	0.95	0.96
400	1.00	0.98	0.99	0.98
1000	1.00	1.00	1.00	1.00
2000	0.99	1.00	1.00	1.00

T	Corresponding values for R^2, averaged over d			
50	0.714	0.731	0.682	0.745
100	0.821	0.781	0.769	0.773
200	0.896	0.857	0.857	0.853
400	0.913	0.901	0.891	0.897
1000	0.905	0.946	0.935	0.944
2000	0.895	0.970	0.961	0.968

Table 3: The influence of the maximum number of iterations, I, when the number of observations, T, is fixed

4.2.3 Joint Convergence

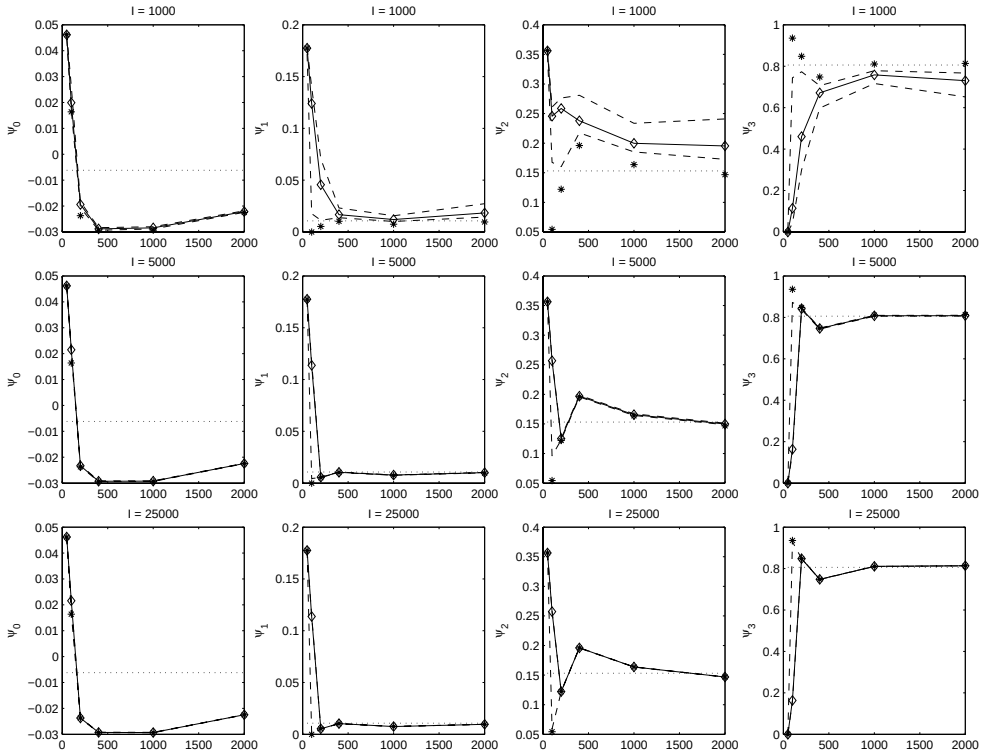
The empirical results on the convergence of the maximum likelihood estimator (subsection 4.2.1), in particular the estimated rate of convergence (Table 2)) confirm the asymptotic theory. Furthermore, when I is chosen at least proportional to T, the threshold accepting approximation of this estimator (subsection 4.2.2) will converge at the same rate to the maximum likelihood estimator. As discussed in Subsection 2.4, the following joint convergence property results: There exists a constant λ_I such that if I is chosen to be $\lambda_I T$, the threshold accepting approximation $\psi_p^{d,T,I,r}$ to ψ_p^{TR} satisfies the convergence condition (5) for any given probability $1-\delta$ and any $\varepsilon > 0$ when T grows at a rate proportional to $1/\varepsilon^2$. Or, expressed in other words, the threshold accepting based maximum likelihood estimator of the GARCH-model parameters converges in probability to the parameters of the data generating process at the same rate as the theoretical maximum likelihood estimator.

Thus, the additional stochastic component introduced by the use of a stochastic search heuristic does not destroy the convergence properties of the maximum likelihood estimator. In fact, as a comparison with a standard approach will show in the following subsection, the heuristic provides better and more robust results.

Figure 4 depicts the convergence of the maximum likelihood estimates to the true parameters in dependence of the number of observations for the considered specimen data series as well as the median and the 10% and 90% quantiles of the results reported by the threshold accepting algorithm for different numbers of iterations. The graphs for $I = 1000$ iterations (top row) exhibit the difficulties of the algorithm to find the optimal results, i.e. the maximum likelihood estimates, when I is chosen too small. When considering the quantiles, it turns out that some of these "bad" reported solutions are actually closer to the true parameters than to the maximum likelihood estimates. However, this might be rather a special feature of the data generating process considered in our paper than a general outcome. Increasing I slightly (second and third row) results in high quality approximations to the maximum likelihood estimator already for rather small sample sizes.

When the sample size is very short, the optimization problem might become more demanding. Thus, the threshold accepting algorithm has a higher chance of converging prematurely to – and eventually reporting a local optimum. In this case, the quality of the results will benefit from an increase in the number of iterations, but only to a limited extent: Once the algorithm has converged to a local optimum, additional iterations will not always drive the search process away from this solution. Therefore, the approach practiced in our application to use

Figure 4: Convergence of maximum likelihood estimators ($*$) to the true parameters (horizontal dotted lines) and median ($-\diamond-$) and 10% and 90% quantiles (dashed lines) for I = 1000, 5000 and 25000 maximum iterations for one specimen data series



several restarts R for a given number of iterations I instead of spending all CPU time on a single run appears to be adequate, in particular for small sample sizes.[8]

4.3 Comparison to Standard Software

In addition to the threshold accepting approach, we also estimated the GARCH parameters using the GARCH package for Matlab; Brooks *et al.* (2001) found this package to be more reliable than several other standard econometric software programmes. This package uses a deterministic approach, repeated runs on the same data series will therefore report equal results. For the 100 different realizations each with six different values for T, 600 different parameter sets are estimated with this package. Comparing the reported value of the loglikelihood function to a ten digit precision, in 253 out of these 600 cases the threshold accepting approach finds better results, in 259 cases the differences of the reported results are below precision, and in just 88 cases, the best of all solutions reported by threshold accepting is inferior to Matlab's. The advantage of the threshold accepting over the deterministic approach becomes even more apparent when the magnitude of the deviations is considered: defining the deviation $\Delta = L^{TA} - L^{\text{Matlab}}$, the "worst" deviation for the threshold accepting is $\Delta = -0.000001$. For the evaluations with respect to the maximum likelihood estimator, these differences are negligible. The "best" result for threshold accepting, however, comes with $\Delta = +9.45$, corresponding to a real desaster of the conventional algorithm.

5 Conclusions

For estimation tasks being slightly more complex than ordinary least squares regression, deterministic algorithms will not always provide the theoretical estimator. In this case, the use of optimization heuristics might be an adequate solution. However, the stochastic features of these algorithms introduce an additional source of uncertainty to the estimator. In addition to the deviation of the theoretical estimator from the parameters of the data generating process due to the finite sample size, the approximation error of the search heuristic has to be taken into account.

If the search heuristic converges to the theoretical estimator for the number of iterations going to infinity, it is possible to derive a joint convergence result. We

[8] For a detailed analysis of the tradeoff between restarts and number of iterations for given computational resources see Winker (2005).

introduce such a convergence result for threshold accepting applied to maximum likelihood estimation. Unfortunately, so far, no distributional results are available for the approximation by threshold accepting. Nevertheless, convergence in probability is a strong result as compared to standard algorithms.

We apply the method to the maximum likelihood estimation of a GARCH-model and find that the theoretical joint convergence result holds for the application already when setting the number of iterations of the algorithm proportional to the sample size. Thus, the threshold accepting based estimator has superior convergence properties compared to standard approaches. Furthermore, it generates better and more robust results already for small samples.

Overall we conclude that the use of stochastic optimization tools in econometrics is indicated whenever standard tools fail to generate reliable results. Furthermore, an application of these tools might provide benchmarks when the quality of standard methods is unknown. Of course, our empirical results resting on a single data generating process, further evidence is required to assess the robustness of our findings. Furthermore, it would be highly interesting to derive the distribution of the results obtained by the optimization tool instead of a convergence in probability result. These extensions are left for future research.

References

Althöfer, I. and K.-U. Koschnik (1991). On the convergence of threshold accepting. *Applied Mathematics and Optimization* **24**, 183–195.

Bollerslev, Tim and Eric Ghysels (1996). Periodic autoregressive conditional heteroscedasticity. *Journal of Business and Economic Statistics* **14**(2), 139–151.

Brooks, Chris, Simon P. Burke and Gita Persand (2001). Benchmark and the accuracy of GARCH model estimation. *International Journal of Forecasting* **17**, 45–56.

Clements, M. P. and H.-M. Krolzig (1998). A comparison of the forecast performance of Markov-switching and threshold autoregressive models. *Econometrics Journal* **1**, C47–C75.

Dorsey, B. and W. J. Mayer (1995). Genetic algorithms for estimation problems with multiple optima, nondifferentiability and other irregular features. *Journal of Business and Economic Statistics* **13**, 53–66.

Dueck, G. and T. Scheuer (1990). Threshold accepting: A general purpose algorithm appearing superior to simulated annealing. *Journal of Computational Physics* **90**, 161–175.

Fitzenberger, B. (1997). A guide to censored quantile regressions. In: *Handbook of Statistics, Volume 15: Robust Inference* (G. S. Maddala and C. R. Rao, Eds.). pp. 405–437. Elsevier. Amsterdam.

Fitzenberger, Bernd and Peter Winker (1998). Using threshold accepting to improve the computation of censored quantile regression. In: *COMPSTAT 1998, Proceedings in Computational Statistics* (Roger Payne and Peter Green, Eds.). Physica. Heidelberg. pp. 311–316.

Herwartz, H. (2004). Conditional heteroskedasticity. In: *Applied Time Series Econometrics* (H. Lütkepohl and M. Krätzig, Eds.). pp. 197–221. Cambridge University Press. Cambridge.

Ho, M. S. and B. E. Sørensen (1996). Finding cointegration rank in high dimensional systems using the Johansen test: An illustration using data based Monte Carlo simulations. *The Review of Economics and Statistics* **78**(4), 726–732.

Maringer, D. (2005). *Portfolio Management with Heuristic Optimization*. Springer. Berlin.

McCullough, B. D. and H. D. Vinod (1999). The numerical reliability of econometric software. *Journal of Economic Literature* **38**, 633–665.

McCullough, B.D. and Berry Wilson (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics & Data Analysis* **31**(1), 27–37.

McCullough, B.D. and Berry Wilson (2002). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. *Computational Statistics & Data Analysis* **40**(4), 713–721.

McCullough, B.D. and Berry Wilson (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics & Data Analysis* **49**(4), 1244–1252.

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

Winker, P. (2005). The stochastics of threshold accepting: Analysis of an application to the uniform design problem. p. forthcoming.