LOCAL LIKELIHOOD ESTIMATORS IN A REGRESSION MODEL FOR STOCK RETURNS[1]

UWE CHRISTIAN JÖNCK

Department Mathematik (SPST), Universität Hamburg
Bundesstr. 55, 20146 Hamburg, Germany

30 March 2007

Abstract

We consider a non-stationary regression type model for stock returns in which the innovations are described by four-parameter distributions and the parameters are assumed to be smooth, deterministic functions of time. Incorporating also normal distributions for modelling the innovations, our model is capable of adapting to light-tailed innovations as well as to heavy-tailed ones. Thus, it turns out to be a very flexible approach. Both, for the fitting of the model and for forecasting the distributions of future returns, we use local likelihood methods for estimation of the parameters. We apply our model to the S&P 500 return series, observed over a period of twelve years. We show that it fits these data quite well and that it yields reasonable one-day-ahead forecasts.

AMS 2000 classifications: primary: 62P20; secondary: 62G08, 91B70, 91B84

Key words and phrases: distributional forecasts, financial time series, local likelihood estimation, local stationarity, Monte Carlo test, non-stationary time series, volatility

Short title: Local likelihood modelling

1 Introduction

During the past twenty years remarkable efforts were made to develop and investigate models for time series of stock returns, i.e. for time series $X_t = \log(P_t/P_{t-1})$, $t = 1, \dots, n$, where P_t represents the price of a stock (or a stock index) at time t. In doing so the ARCH and GARCH models of Engle (1982) and Bollerslev (1986) were in the centre of interest. The basic GARCH(p, q) model takes the form

$$(1.1) \qquad\qquad X_t = \sigma_t \varepsilon_t,$$

with white noise $\{\varepsilon_t\}$ and volatilities given by

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2.$$

There have been various modifications on this approach, which led to more sophisticated but on the other hand also more complicated models. For an overview see for example Bollerslev

---

et al. (1994) or Gouriéroux (1997). These models succeed in reproducing various stylized facts. However, especially when dealing with long time series, they also have some drawbacks, such as the occurence of the so-called IGARCH effect in the popular GARCH(1,1) model. (A critical discussion of this model can be found in Stărică, 2003.) Mikosch and Stărică (2003, 2004) argue that this effect could be due to the fact that time series of log returns are actually not stationary, which is one of the basic paradigms of the GARCH class models. Furthermore, they show that also the so-called long range dependence effects in the second moment structure can be explained if the underlying time series is non-stationary. The assumption of stationarity might also seem questionable especially for long financial time series, because the financial markets and the general economic framework may change over long time periods. Consequently, the concept of stationarity might not be adequate for a financial time series model. A weakening of the strong requirement of stationarity might be the concept of a *locally stationary* behaviour of the time series: this means that for small time intervals the time series is nearly stationary, whereas over longer time intervals its (stochastic) behaviour changes gradually.

Drees and Stărică (2002) propose a very simple regression-based model that drops the assumption of stationarity. For the returns X_t Drees and Stărică presume a multiplicative structure, similar to that of (1.1). The difference to the GARCH models lies in the modelling approach for the volatilities σ_t. In the GARCH class models these are determined by endogenous factors, namely by past volatilities as well as past returns. In contrast to that, Drees and Stărică assume the volatilities to be driven by some unknown (presumably rather complex) market conditions, and hence to be determined exogenously. However, this exterior mechanism might be too complex to develop a realistic stochastic model for it. For this reason Drees and Stărică simply model the volatilities as a smooth, deterministic function of time. Thus, their basic model is the following:

$$(1.2) \quad \begin{cases} X_t = \mu + \sigma(t)\varepsilon_t, \quad t = 1, 2, \ldots, n, \quad \mu = \text{const.}, \\ \{\varepsilon_t : t = 1, \ldots, n\} \sim \text{IID}(0, 1), \\ \sigma(t), \ t = 1, 2, \ldots, n, \quad \text{a smooth, deterministic function of time.} \end{cases}$$

Considering the centered returns $R_t = X_t - \mu$, we have

$$R_t^2 = \sigma(t)^2 + \sigma(t)^2(\varepsilon_t^2 - 1).$$

Hence, (1.2) can be interpreted as a regression model for the squared volatilities, and non-parametric regression techniques can be applied for estimating the volatilities. The innovations ε_t are modelled by asymmetric Pearson Type VII distributions, which generalize the class of t-distributions (see below), and thus allow for heavy tails. Drees and Stărică fit this model to the daily returns of the closing prices of the Standard & Poor's 500 (S&P 500) stock index from 2 January 1990 to 21 February 2002, depicted in Figure 1.1. They estimate the volatilities by a simple Nadaraya-Watson kernel estimator and they use a maximum likelihood (ML) estimator to determine the parameters of the t-distributions. They show that their model is competitive to the conventional GARCH(1,1) and EGARCH(1,1) models concerning the fitting of the data as well as the forecasting of future return distributions.

In this paper (which can be regarded as a continuation of the work of Drees and Stărică, 2002) we take up and refine the above regression model. However, instead of the assump-
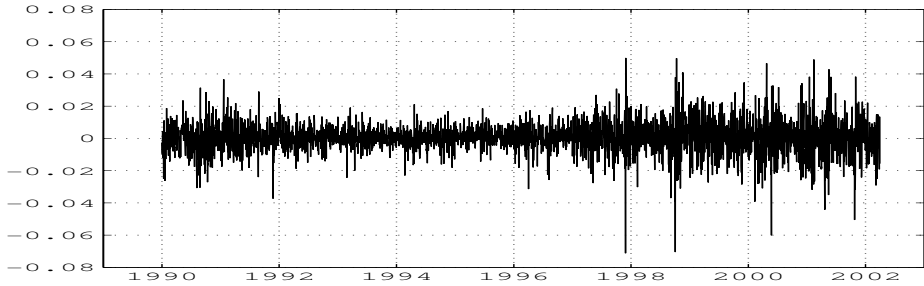
Figure 1.1. *Daily returns of the closing prices of the S&P 500 index from 2 January 1990 to 21 February 2002.*

tion of identically distributed innovations, we model the parameters of the distributions as smooth functions of time. Furthermore, we expand the model by modelling innovations with very light tails by normal distributions. These two new aspects increase the model's flexibility. For both, the fitting of this refined model to the S&P 500 return series and the forecasting of future return distributions, we use local likelihood estimators. These estimators combine the idea of ML estimation and nonparametric regression techniques. We discuss their asymptotic behaviour, which enables us to construct asymptotic confidence intervals for the true parameters. The theoretical background for these asymptotics is based on a paper of Aerts and Claeskens (1997). Assessments of the estimated parameters and the estimated forecasts, respectively, show that our approach seems to provide a good fitting of the data as well as reasonable forecasts.

The paper is organized as follows: in Section 2 we present the underlying model. In Section 3 we give a brief overview of the method of local likelihood estimation, and we discuss the construction of asymptotic confidence intervals. In Section 4 we fit the model to the S&P 500 data and evaluate the results with several testing devices. We also provide forecasts of future return distributions, which is done in Section 5. Section 6 concludes.

2   Refinement of the model of Drees and Stărică

We consider the regression model (1.2) from the previous section. In the following, instead of the returns X_t we consider the centered returns $R_t = X_t - \mu$. Then we have (approximately)

$$(2.1) \qquad\qquad R_t = \sigma(t)\varepsilon_t \approx X_t - \overline{X}_n,$$

regarding $\overline{X}_n = 1/n \sum_{j=1}^{n} X_j$ as an approximation for $\mu = \mathbb{E}X_t$, and neglecting an estimation error that might occur. We assume the innovations ε_t—and, consequently, the centered returns, too—to be independent and distributed with density

$$\mathbf{g}(x; u_+, u_-, v_+, v_-) = \frac{1}{2}\Big(g(|x|; u_-, v_-)\mathbb{1}\{x < 0\} + g(x; u_+, v_+)\mathbb{1}\{x \geq 0\}\Big),$$
$$u_\pm \in [0, 2), \quad \tau_\pm \in (0, \infty),$$

where

$$(2.2) \qquad g(x; u, v) = \begin{cases} \frac{2\Gamma(1/u)}{\Gamma(1/u - 1/2)\sqrt{\pi}} \cdot \frac{\sqrt{u}}{v} \left(1 + \left(\frac{x}{v}\right)^2 u\right)^{-1/u} \mathbb{1}\{x \geq 0\}, & u > 0, \\ \\ \frac{2}{v\sqrt{\pi}} \exp\left\{-\left(\frac{x}{v}\right)^2\right\} \mathbb{1}\{x \geq 0\}, & u = 0. \end{cases}$$

The corresponding cdf is denoted $\mathbf{G}(\cdot; u_+, u_-, v_+, v_-)$, or sometimes just $\mathbf{G}$. The parameters u_+ and u_- essentially determine the shape of the density $\mathbf{g}$: for $u > 0$ the term $1/2g(\cdot; u, v)$ equals the right tail of the density of a t-distribution with $2/u - 1$ degrees of freedom and scale parameter $v/(2 - u)^{1/2}$ (which is sometimes also referred to as a Pearson Type VII distribution, cf. Johnson and Kotz, 1970, p. 114ff.), and for $u = 0$ it equals the right tail of the density of a normal distribution with mean zero and variance $v^2/2$. The parameters v_+ and v_- behave like scale parameters for the positive and negative parts of the distribution, respectively. It is easily seen that $\mathbf{G}$ has finite variance iff $u_+, u_- < 2/3$. This modelling approach is more flexible than that of Drees and Stărică (2002), who always assume $u_+, u_- > 0$. In contrast, our model can adapt to heavy-tailed as well as light-tailed innovations—as we shall see later on.

Note that for each x we have $g(x; u, v) \to g(x; 0, v)$ as $u \to 0$. Hence, $g(x; 0, v)$ can also be considered an approximation to $g(x; u, v)$ for small values u. In the following this property is often referred to as the "normal approximation" of the tails.

Another nice property of the law $\mathbf{G}$ is that it allows for asymmetry in the tails, and thus is quite flexible. It may appear somewhat inconsequent to model the innovations (and thus the centered returns)—presumed to have mean zero—by a class of distributions that in general do not have this property (the distributions $\mathbf{G}$ are in general not centered, but only have median zero). However, for practical applications this is not of great importance: indeed, the model assessments in Section 4 will show that the means of the fitted distributions are approximately equal to zero.

We further refine the model of Drees and Stărică by dropping their assumption of identically distributed innovations. Instead, we assume the parameters to be smooth, deterministic functions of time, i.e. we have as an initial approach

$$\varepsilon_t \sim \mathbf{G}(\cdot; u_+(t), u_-(t), v_+(t), v_-(t)).$$

To avoid an overparametrization resulting from the additional scale parameter $\sigma(t)$, we introduce new parameters $\tau_+(t) := \sigma(t)v_+(t)$ and $\tau_-(t) := \sigma(t)v_-(t)$, which leads to the modelling assumption of independent centered returns

$$(2.3) \qquad R_t \sim \mathbf{G}(\cdot; u_+(t), u_-(t), \tau_+(t), \tau_-(t)).$$

Hence, our model is completely determined by the distributions of the centerd returns, given by (2.3), and the smooth parameter function

$$\theta : \{1, \ldots, n\} \to [0, 2)^2 \times (0, \infty)^2, \quad t \mapsto \begin{pmatrix} \theta_1(t) \\ \theta_2(t) \\ \theta_3(t) \\ \theta_4(t) \end{pmatrix} \equiv \begin{pmatrix} u_+(t) \\ u_-(t) \\ \tau_+(t) \\ \tau_-(t) \end{pmatrix}.$$

In the following, we shortly write $\mathbf{G}(\cdot;\theta)$ or $\mathbf{G}_\theta$, instead of $\mathbf{G}(\cdot;u_+,u_-,\tau_+,\tau_-)$, and for the corresponding density we write $\mathbf{g}_\theta$. Note that in this approach the volatilities only play a subordinate role as part of the parameters τ_+ and τ_-. However, assuming that for the innovations we have $\mathrm{Var}(\varepsilon_t) = 1$, the volatilities can easily be regained from the equation $\sigma^2(t) = \mathrm{Var}(R_t)$.

3   Local likelihood estimation

For fitting our model to data we use the method of local likelihood estimation. In this section, we first give a brief overview of its basic concept. Subsequently, based on a paper by Aerts and Claeskens (1997), we investigate the asymptotics of these estimators, and we provide a device for constructing approximate confidence intervals for the parameters. We discuss how these ideas carry over to our model.

The basic concept. Consider a sample $(x_1, y_1), \ldots, (x_n, y_n)$. The data x_i denote distinct points of time in some interval $[a, b]$. These might either be fixed or given by a random design. The corresponding y_i's are realizations of independent, real-valued random variables Y_i. Suppose that each Y_i is distributed according to some pdf $f(\cdot, s_i)$, where the parameter $s_i \in \mathbb{R}^d$ is determined by a function $\theta\colon x_i \mapsto \theta(x_i) = s_i$. Assuming this parameter function θ to be sufficiently smooth, each of its components θ_k can be approximated locally by a Taylor polynomial of degree p_k. Thus, for x_i close to a fixed point x we have the approximation

$$(3.1) \qquad \theta_k(x_i) \approx \sum_{j=0}^{p_k} \frac{\theta_k^{(j)}(x)}{j!}(x_i - x)^j \equiv \sum_{j=0}^{p_k} \beta_{kj}(x_i - x)^j,$$

where in this notation the dependence of β_{kj} from x is omitted. For estimating the coefficients β_{kj} of the local polynomials, we consider the *local likelihood function*

$$(3.2) \qquad \begin{cases} L_n(\beta; h, x) = \displaystyle\sum_{i=1}^n \log f\left(y_i, \sum_{j=0}^{p_1}\beta_{1j}(x_i - x)^j, \ldots, \sum_{j=0}^{p_d}\beta_{dj}(x_i - x)^j\right) K_h(x_i - x), \\ \beta = (\beta_1^T, \ldots, \beta_d^T)^T = (\beta_{11}, \ldots, \beta_{1p_1}, \ldots, \beta_{d1}, \ldots, \beta_{dp_d})^T. \end{cases}$$

Here K denotes a kernel, h is a bandwidth (or smoothing parameter), and $K_h(\cdot) = K(\cdot/h)/h$. Usually K is a unimodal symmetric pdf, preferably with compact support. Maximizing $L_n(\beta; h, x)$ with respect to β yields the local likelihood estimates $\hat{\beta}_{kj}$. Consequently, we have estimates $\hat{\theta}_k^{(j)}(x) = j!\hat{\beta}_{kj}$ for the value of the parameter function in x and its derivatives, respectively. For further information on local likelihood estimation and local polynomial fitting cf. Aerts and Claeskens (1997), Fan et al. (1998), or Fan and Gijbels (1996).

Asymptotics. In the following, we assume the design points x_i to be generated according to $x_i = G^{-1}\big((i-1)/(n-1)\big)$, with $G(x) = \int_{-\infty}^x f_X(t)dt$ and a design density f_X with $\mathrm{supp}(f_X) = [a, b]$. Furthermore, we assume the kernel K to be a symmetric pdf with compact support $[-1, 1]$. Aerts and Claeskens (1997) show that under some additional regularity conditions the local likelihood estimators are consistent and asymptotically normal. (Mainly, these regularity conditions are the classical conditions on the densities $f(\cdot, \theta)$, needed for

proving consistency and asymptotic normality of ML estimators; cf. Aerts and Claeskens, 1997, conditions (R1)–(R6), for details.) More precisely, they show that

$$(3.3) \quad \sqrt{nh} \left\{ \left((\hat{\beta}_1(x) - \beta_1^{\star}(x))^T \mathbf{H}_{p_1}, \ldots, (\hat{\beta}_d(x) - \beta_d^{\star}(x))^T \mathbf{H}_{p_d} \right)^T - \frac{1}{\sqrt{nh}} \mathbf{B}_{nh}(x) \right\}$$
$$\xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{V}(x)),$$

as $n \to \infty$, with $\beta_k^{\star}(x) = (\beta_{k1}^{\star}(x), \ldots, \beta_{kp_k}^{\star}(x))^T$ denoting the vector of the true coefficients of the local polynomials, i.e. $\beta_{kj}^{\star}(x) = \theta_k^{(j)}(x)/j!$, and employing the indexing given in (3.2). The matrices $\mathbf{H}_{p_k}$ are diagonal matrices with entries $h^0, \ldots, h^{p_k}$, whereas the other quantities depend in a rather complicated way on the true parameters $\theta_k^{(j)}(x)$, the kernel K, the (log-)densities and its derivatives and on the design point x. Again, cf. Aerts and Claeskens (1997) for details.

Confidence intervals. The results from the previous paragraph provide a device for constructing approximate pointwise confidence intervals for the true coefficients of the local polynomials. With $\mathbf{e}_{ij}$ denoting the vector with "1" as ij-entry and "0" elsewhere (i.e., we have $\mathbf{e}_{ij}^T \beta = \beta_{ij}$), from (3.3) we get

$$\sqrt{nh} \left\{ (\hat{\beta}_{ij}(x) - \beta_{ij}^{\star}(x)) h^j - \frac{1}{\sqrt{nh}} \mathbf{e}_{ij}^T \mathbf{B}_{nh}(x) \right\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{e}_{ij}^T \mathbf{V}(x) \mathbf{e}_{ij}).$$

Following the argumentation of Fan and Yao (2003), p. 243, we neglect the asymptotic bias term. Hence, approximately, we have

$$\mathscr{L}(\hat{\beta}_{i0}(x) - \beta_{i0}^{\star}(x)) \approx \mathcal{N}\left(0, \frac{1}{nh} \mathbf{e}_{i0}^T \mathbf{V}(x) \mathbf{e}_{i0}\right).$$

Replacing the parameters $\theta_i(x)$ in the expression for the asymptotic covariance matrix $\mathbf{V}(x)$ by the corresponding estimates $\hat{\theta}_i(x)$ we get an approximate covariance matrix $\widehat{\mathbf{V}}(x)$. Thus, by

$$\left[\hat{\beta}_{i0}(x) - u_{\alpha/2} \left(\widehat{\mathbf{V}}_{i0,i0}(x)/(nh) \right)^{1/2} \; ; \; \hat{\beta}_{i0}(x) + u_{\alpha/2} \left(\widehat{\mathbf{V}}_{i0,i0}(x)/(nh) \right)^{1/2} \right]$$

we get an approximate confidence interval at the level $1 - \alpha$ for the true parameter $\theta_i(x)$, where $u_{\alpha/2}$ is the standard normal $(1 - \alpha/2)$-quantile.

Bandwidth selection. A crucial point in local likelihood estimation is the choice of an appropriate bandwidth h. This acts as a smoothing parameter in nearly the same way as it does in ordinary nonparametric regression. The above asymptotics motivate to chose a bandwidth that minimizes the asymptotic mean squared error of the estimates, which is a function of h and also depends on some unknown constants. The latter ones have to be estimated by plug-in procedures. (Cf. also Aerts and Claeskens, 1997.) Another approach, which is not based on the above asymptotics, is discussed by Fan et al. (1998): they derive estimates for the bias and variance terms, and thus for the mean squared error (MSE) of the local likelihood fit. Following their approach the bandwidth is chosen such that the estimated MSE, which is a function of h, is minimal. This method is especially feasible for the computation of local bandwidths. A nice alternative to the above bandwidth selection procedures is a cross-validation method, proposed by Aerts and Claeskens (1997). In this fully data-driven approach for a given bandwidth h the parameters $\theta(x_1), \ldots, \theta(x_n)$

are estimated, where the ith estimate is based on the sample without the ith oberservation (x_i, y_i). The resulting parameter estimates—denoted $\hat{\theta}^{[i]}(x_i)$—are then plugged into the log-likelihood function. Doing these computations for varying bandwidths h yields the "cross-validation" function

$$CV(h) = \sum_{i=1}^{n} \log f(y_i; \hat{\theta}^{[i]}(x_i)),$$

which is then maximized with respect to h. This cross-validation approach is easy to implement and we will also consider it later for our estimation problems in Sections 4 and 5.

Applying the asymptotics to our model.   For applying the above asymptotics to our modelling approach, we have to consider some minor modifications to our model. After centering the returns (cf. (2.1)), we are given observations $\{(t, R_t) : t = 1, \ldots, n\}$. For the purpose of asymptotic investigations, now we consider a rescaling of the time points by the transformation

(3.4) $$s \mapsto \frac{s-1}{n-1} =: \tilde{s}, \quad 1 \leq s \leq n,$$

which maps the time points into the unit interval. These rescaled design points fulfill the assumption on the points x_i from the above paragraph on asymptotics with $f_X(x) = \mathbb{1}_{[0,1]}(x)$. The parameter function can be rescaled in the same way, setting $\tilde{\theta}(\tilde{t}) := \theta(t)$. Note that this transformation is only necessary for the application of the asymptotic results and the computation of the confidence intervals. It has no effects on practical aspects of the estimation. Therefore, when it is obvious that an assertion is correct for the "original" points of time (i.e. for $t = 1, \ldots, n$) as well as for the rescaled ones (i.e. for the points $\tilde{t}$), for ease of notation we often silently omit the notation for the rescaling. Especially, we do so when considering the parameter function and the local polynomials.

In addition, we have to check whether the regularity conditions on the densities hold for our model. Actually, this is the case for a suitable submodel, namely for the family $\{\mathbf{G}_\theta : \theta \in \Theta\}$. Here the parameter space is defined as $\Theta := (\underline{u}, \overline{u})^2 \times (\underline{\tau}, \overline{\tau})^2$, with arbitrary values $0 < \underline{u} < \overline{u} < 2/3$ and $0 < \underline{\tau} < \overline{\tau} < \infty$. The verification of the regularity conditions is lengthy but quite standard. It is based upon showing diverse interchangeability conditions for integration and differentiation of the densities and rough estimation inequalities for the derivatives of the (log-)densities; it also uses the fact that $\overline{\Theta}$ is compact. The complete proof is given in Jönck (2005). Consequently, for estimations within this submodel we can calculate approximate confidence intervals for the true parameters according to Section 3. Since $\theta \mapsto (\operatorname{Var} \mathbf{G}_\theta)^{1/2}$ is differentiable in θ, an application of the Delta method yields approximate confidence intervals for the volatilities, too.

4   Modelling the S&P 500 return series

In this section we fit our model to the returns of the closing prices of the S&P 500 return series from 2 January 1990 to 21 February 2002 (cf. Figure 1.1), which is a total of $n = 3062$ observations. We discuss the numerical results and check whether they fulfill the model assumptions.

Fitting the model. We estimate the (rescaled) parameter function $\theta(x)$ for each of the points $x = \tilde{x}_1, \ldots, \tilde{x}_n$, with $\tilde{x}_i = (i-1)/(n-1)$, cf. equation (3.4). To this end, for each such x the optimization problem

$$
(4.1) \quad
\begin{cases}
\max_{\beta} L_n(\beta; h, x) \quad \text{under constraints:} \quad (\beta_{10}(x), \ldots, \beta_{40}(x))^T \in \Theta, \\[2mm]
\left(\sum_{j=0}^{p_1} \beta_{1j}(\tilde{x}_i - x)^j, \ldots, \sum_{j=0}^{p_4} \beta_{4j}(\tilde{x}_i - x)^j \right)^T \in \Theta, \quad \tilde{x}_i \in [x-h, x+h],
\end{cases}
$$

is solved, where $L_n(\beta; h, x)$ is the local likelihood function defined by (3.2), employing the densities $\{\mathbf{g}_\theta : \theta \in \Theta\}$. For the kernel K we use the Epanechnikov kernel, which is given by $K(u) = 3/4(1-u^2)^+$. Using the notation from (3.2), the vector $(\beta_{10}(x), \ldots, \beta_{40}(x))^T$ corresponds to $\theta(x) = (u_+(x), u_-(x), \tau_+(x), \tau_-(x))^T$. For the computations we use the `fmincon` routine implemented in `Matlab` (V.7). The meaning of the first constraint in (4.1) is obvious, the second one ensures that all evaluations of the local polynomials lie in the parameter space, otherwise $L_n(\beta; h, x)$ is not defined. We take Θ as defined in Section 3, with $\underline{u} = 10^{-2}$, $\overline{u} = 2/3$, $\underline{\tau} = 10^{-4}$ and $\overline{\tau} = 5$.

We compute the cross-validation function $CV(\cdot)$ for bandwidths ranging from 0.010 to 0.075 (cf. Section 3): for bandwidths between 0.020 and 0.050 it is very flat, and it has a (not very distinct) maximum at $h_{CV} = 0.026$. For our computations we choose the bandwidth $h = 0.030$.

We use local linear fits for all four paramters, i.e. $p_1 = \cdots = p_4 = 1$. It is preferable to use the same degree of adaption for all parameters, since the expression for the asymptotic bias is essentially of an order of the smallest such polynomial degree. In addition, the use of odd-degree polynomial fits reduces boundary effects. (For details see Aerts and Claeskens, 1997.) Our choice of local linear fits takes account of both aspects and at the same time keeps the number of optimization parameters to a minimum.

In addition, we use the normal approximation from Section 2 in the following way: if for some x the local likelihood estimation yields an estimate $\hat{u}_+(x) = \underline{u}$, we can assume the true value $u_+(x)$ to be very close to zero, thus implying a very light right tail of the corresponding return distribution $\mathscr{L}(R_t)$. According to the above ideas, then we can approximate the right tail of the return distribution by that of a $\mathscr{N}(0, \tau_+^2/2)$-distribution, and thus $\hat{u}_+(x)$ is assigned the value zero. Likewise we proceed if $\hat{u}_-(x) = \underline{u}$. In case of either $\hat{u}_+(x)$ or $\hat{u}_-(x)$ being zero, the conditions for the asymptotic results by Aerts and Claeskens (1997) are violated. Therefore, in such cases we do not provide asymptotic confidence intervals for any of the four paramters $u_+(x)$, $u_-(x)$, $\tau_+(x)$ and $\tau_-(x)$.

Numerical results. The estimates for the parameter functions are shown in Figure 4.1. Apparently, the normal tail approximation is widely employed, and so it seems a valuable tool to enhance the model's ability of assessing light-tailed distributions.

One can clearly see that for most points of time we have $\hat{u}_+ < \hat{u}_-$, i.e. the left tails of the estimated distributions are heavier than the corresponding right tails. This might reflect the well-known fact that in time series of stock returns the extremely large negative returns are usually larger in absolute value than the extremely large positive returns. Furthermore, the estimated distributions show some sort of symmetric behaviour in the tails. For example,
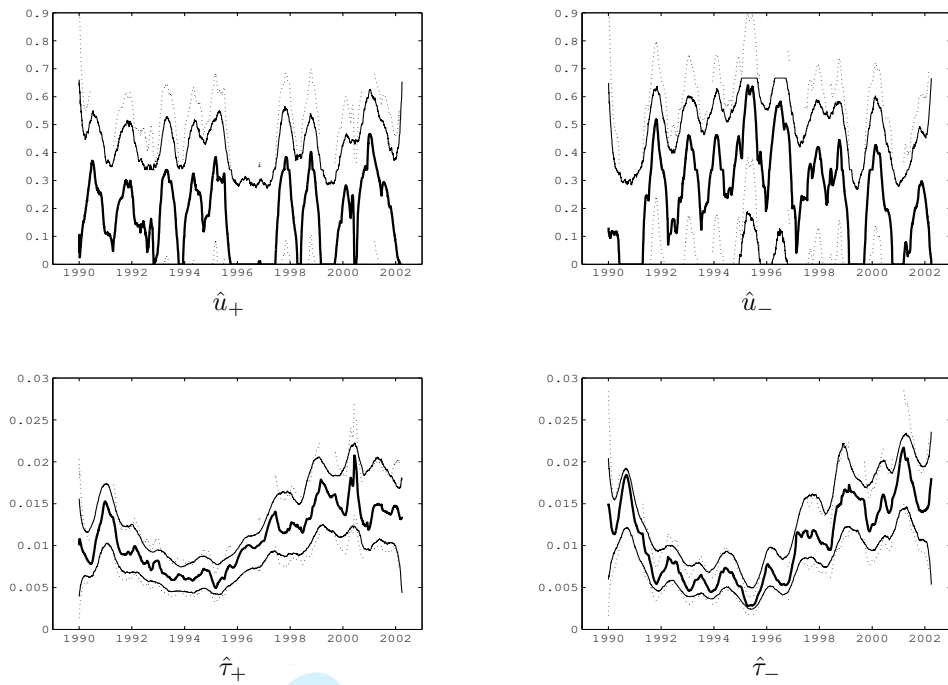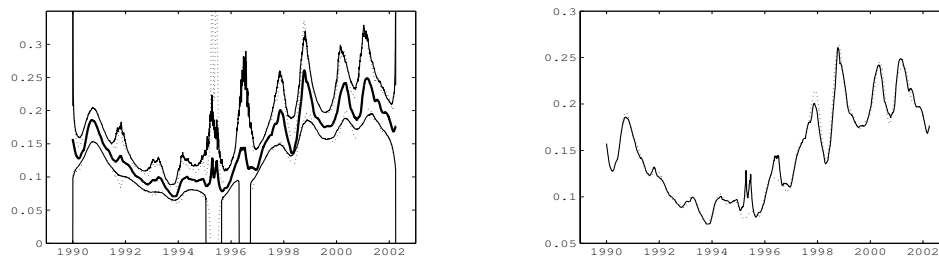
Figure 4.1. *The estimated parameter functions $\hat{u}_+$, $\hat{u}_-$, $\hat{\tau}_+$ and $\hat{\tau}_-$ (bold solid lines) together with the corresponding asymptotic pointwise 95%-confidence intervals (dotted lines), where applicable, and the bootstrap confidence intervals (thin solid lines).*

whenever the left tails of the distributions get heavier, i.e. $\hat{u}_-$ increases, the parameter function $\hat{u}_+$ also tends to increase and vice versa. Apparently, such a symmetric behaviour can be observed in the estimated scale parameters $\hat{\tau}_+$ and $\hat{\tau}_-$, too, where this symmetry seems even much more distinctive. Remember that we defined $\tau_\pm(t) = \sigma(t)v_\pm(t)$. Thus, due to the symmetric behaviour of the scale parameters one might try to simplify the model in such a way that only the volatility $\sigma(t)$ is presumed a smooth function and the parameters $v_+(t)$ and $v_-(t)$ are presumed global constants. However, it turns out that the quotient $\hat{\tau}_+(t)/\hat{\tau}_-(t)$ is strongly fluctuating with varying t and that the confidence intervals for $\tau_+(t)/\tau_-(t)$ (which can be derived easily from the use of the Delta method) are nearly disjoint for many points of time.

The estimates $\hat{u}_+$ and $\hat{u}_-$ also seem to exhibit a certain kind of periodicity, which expresses through the fact that the periods between two consecutive points of maximum (and minimum, respectively) are approximately of the same length. This effect might be due to the choice of the rather small bandwidth h—for the estimation of the parameters at any point of time less than 185 observations are taken into account. Indeed, for larger bandwidths this "periodicity" appears much less distinct.

Figure 4.1 also shows the pointwise confidence intervals for the true parameters, as derived in Section 3 (depicted in dotted lines). Since we can only provide such asymptotic confidence intervals for those points of time for which no use of the normal approximation is made for either tail (see above), we also compute 95%-bootstrap confidence intervals for each of the estimated parameters (depicted in thin solid lines in Figure 4.1). To this end, we simulate a total of $J = 1000$ time series samples $\{R_t^{*j} : t = 1, \ldots, n\}$, with independent

Figure 4.2. *Left: The annualized volatilities $\hat{\sigma}_{\mathrm{a}}$ (bold solid lines) with the corresponding asymptotic pointwise 95%-confidence intervals (dotted lines), where applicable, and the bootstrap confidence intervals (thin solid lines). Right: The estimated annualized volatilities from our local likelihood estimation approach (solid lines) compared to the estimates from Drees and Stărică (2002), which are based on a kernel estimate (dotted lines).*

variables

$$(4.2) \qquad R_t^{*j} \sim \mathbf{G}_{\hat{\theta}(t)}, \quad t = 1, \ldots, n, \quad j = 1, \ldots, J.$$

For each of these simulated samples we compute likelihood estimates $\{\hat{\theta}^{*j}(t) : t = 1, \ldots, n\}$. Thus, for every t and every k (denoting one of the parameter components) the upper and lower 2.5%-quantiles of the set $\{\hat{\theta}_k^{*j}(t) : j = 1, \ldots, J\}$ yield a 95%-boostrap confidence interval for the parameter $\hat{\theta}_k(t)$. This is a very simple procedure for the construction of bootstrap confidence intervals. However, the resulting intervals should be interpreted carefully, because we do not know anything about their (asymptotic) properties and thus their accuracy. In addition, we do not know the effects on these intervals which occur due to the fact that for parameter values u_+ or u_- close to zero our model switches from t-distributions to normal tails. Nevertheless, these bootstrap confidence intervals are helpful to get a first impression on how the local likelihood estimates in our modelling approach behave.

For u_+ and u_- the asymptotic confidence intervals (as well as the bootstrap confidence intervals) are rather wide, in some cases they cover nearly all of the space $(\underline{u}, \overline{u})$. This is somewhat unsatisfying, because it yields little information about the accuracy of our estimated functions for the parameters u_+ and u_-: these confidence intervals would as well allow for parameter functions u_+ and u_- of a different shape, for example they would allow for globally constant parameter functions. It might therefore be worth thinking about a simplified approach, modelling u_+ and u_- as global constants, like Drees and Stărică (2002) do. We shall come back to and further discuss this idea at the end of this section. For the scale parameters τ_+ and τ_- the asymptotic confidence intervals as well as the bootstrap confidence intervals are rather narrow. Indeed, they seem to coincide more ore less.

Figure 4.2 shows the estimated annualized volatilities $\hat{\sigma}_{\mathrm{a}}(t) := \sqrt{250}\hat{\sigma}(t)$, where the $\hat{\sigma}$'s are computed as the standard deviations of the estimated distributions. The corresponding confidence intervals for the estimated volatilities result from the use of the Delta method. The estimated annualized volatilities resemble very much the corresponding estimates from Drees and Stărică, concerning both, the oscillation behaviour and the magnitude of the estimates, as can be seen from the right image in Figure 4.2. Only the double-peak in the middle of 1995 appears a bit strange, in particular, since the time series is very calm during

that period. The occurence of such effects depends of course on the choice of the bandwidth: they can be avoided by chosing a larger bandwidth, which on the other hand might lead to an over-smoothing of some parts of the curve. To completely prevent such problems, the use of a local, data-driven bandwidth (as proposed by Fan et al., 1998, Section 6) may be helpful, which is however computationally very intensive and therefore not always recommended.

Assessment of the model.   Summarizing so far, one can say that the local likelihood estimators in our model, together with the normal tail approximation seem to yield reasonable results for describing the development of the time series. Next we want to check whether the model assumptions are really fulfilled. Of course, this also gives some answer to the question of the goodness of our modelling approach. We have to check whether

- the (centered) returns are independent,

- the estimated distributions $\mathbf{G}_{\hat{\theta}(t)}$ are good approximations to the real distributions of the returns, i.e. approximately we have $R_t \sim \mathbf{G}_{\hat{\theta}(t)}$, and

- the estimated distributions $\mathbf{G}_{\hat{\theta}(t)}$ have approximately mean zero (i.e. the "centered" returns R_t are really centered).

For this purpose we mainly use the same techniques as Drees and Stărică (2002). Furthermore, we check whether

- the asymptotic confidence intervals cover the true parameter with probability 0.95.

First we check whether the innovations or, equivalent, whether the centered returns R_t are independent. This is one of the basic assumptions of our model. As the common tests for independence always demand for a sample of identically distributed observations, we first standardize the centered returns. Assuming the estimated distributions are good approximations for the return distributions, i.e. we have approximately $R_t \sim \mathbf{G}_{\hat{\theta}(t)}$, the variables

$$(4.3) \qquad\qquad V_t := \Phi^{-1}\big(\mathbf{G}(R_t; \hat{\theta}(t))\big), \quad t = 1, \dots, n,$$

—in the following referred to as *standardized returns*—should be standard normal, and thus in case of independence of the returns should be Gaussian white noise. To check this, we investigate the SACFs of the time series of standardized returns and of their absolute values; see Figure 4.3. For almost all lags both of the SACFs stay within the 95%-confidence intervals, given by the dashed lines, thus strongly supporting the independence assumption. In addition, Figure 4.3 shows the *p*-values of the Portmanteau tests for the corresponding time series for the first 100 lags. While for the standardized returns the hypothesis of independence is rejected for almost all lags at the 95%-level, the test supports the hypothesis of uncorrelated absolute standardized returns for almost all lags. (For general information on the confidence intervals or the Portmanteau test cf. Brockwell and Davis, 1987, Section 7.2 and p. 300ff., respectively.) This surprising effect occurs in a similar form also in the original model of Drees and Stărică (2002). To find an explanation for this effect, we consider the SACFs of the returns and the absolute returns of the S&P 500, respectively, which are given in Figure 4.4: the SACF of the absolute returns exhibits some long range dependencies (Fig-
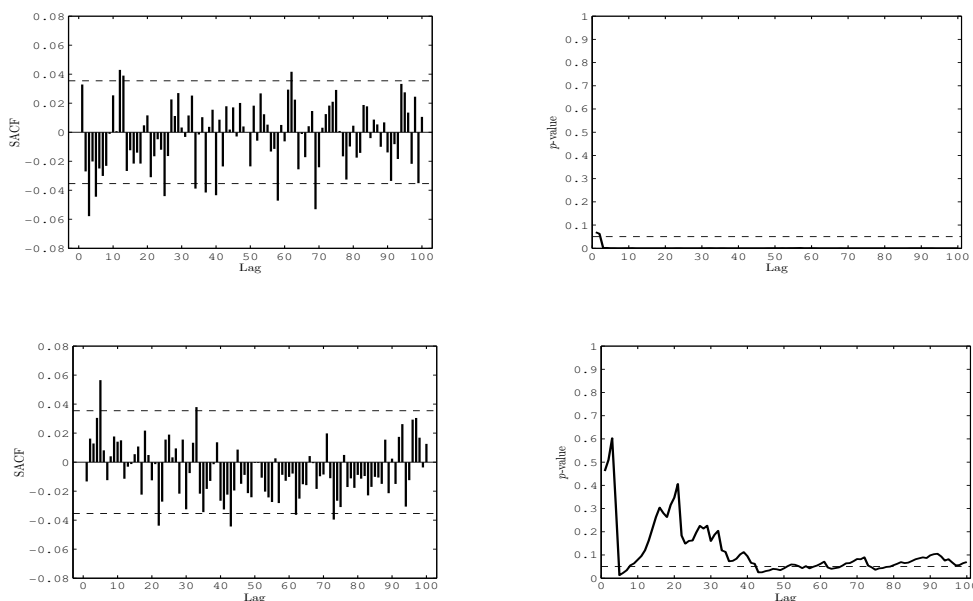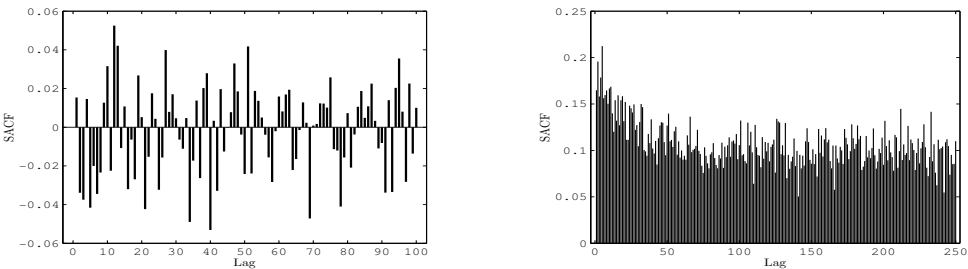
Figure 4.3. *The SACFs (left) and the p-values of the Portmanteau tests (right) for the first 100 lags of the standardized returns $\{V_t : t = 1, \ldots, n\}$ (top) and for the absolute values of the standardized returns $\{|V_t| : t = 1, \ldots, n\}$ (bottom). The dashed lines in the plots on the left represent the 95%-confidence intervals under the hypothesis of independence.*

ure 4.4, right), which, according to Drees and Stărică, might reflect certain non-stationarities within the second moment structure of the time series. They argue that the locally stationary modelling approach describes these non-stationarities quite well. Consequently, the absolute standardized returns behave nearly like uncorrelated random variables, as can be seen by the evaluations of the corresponding SACF and the Portmanteau test (Figure 4.3, bottom). However, this does not necessarily hold true for the standardized returns themselves. Actually, there also might be some non-stationarities (for instance in the mean of the time series), which mainly manifest through the SACF of the returns, and which are not taken into account by our modelling approach. Because the standardization (4.3) itself only changes the absolute values of the returns, but not their signs, the SACFs of both, the returns (Figure 4.4, left) and the standardized returns (Figure 4.3, top left) do not differ very much, and thus our modelling approach cannot capture these effects, which might be a reason for the negative outcome of the corresponding Portmanteau test. Neglecting this flaw, we conclude that the standardized returns (and thus the returns) can—at least approximately—be regarded as independent, in accordance to our model assumptions.

Next we check whether the estimated distributions $\mathbf{G}_{\hat{\theta}(t)}$ are good approximations for the true distributions of the centered returns. That is, we test for the null hypothesis

$$R_t \sim \mathbf{G}_{\hat{\theta}(t)}, \quad t = 1, \ldots, n.$$

If this holds true, then the standardized returns should be approximately standard normal. This is clearly supported by the QQ-plot for the standardized returns, which is given in Figure 4.5. To formally test the normality of the standardized returns, we employ the

Figure 4.4. *The SACFs of the S&P 500 returns (left) and of their absolute values (right) for the first 100 and 250 lags, respectively.*
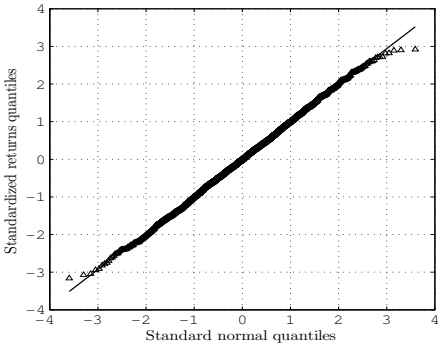
Figure 4.5. *QQ-Plot of the standardized returns $\{V_t : t = 1, \ldots, n\}$.*

Kolmogorov-Smirnov (KS) test and the Jarque-Bera (JB) and Shapiro-Wilk (SW) tests. The latter ones are specially designed for testing normality; cf. Shapiro and Wilk (1965), and Judge et al. (1988), respectively. (The Matlab implementations used for the SW and JB tests are part of the UCSD GARCH Toolbox, available at http://econ.ucsd.edu/~ksheppar; we corrected some errors in the codes.) The high p-values of all three tests, given in the first line of Table 4.1, indicate that the estimated distributions approximate the true return distributions very well.

Finally, we address to the question if the estimated distributions $\mathbf{G}_{\hat{\theta}(t)}$ (approximately) have mean zero, as postulated by the model. If so, according to the central limit theorem, the statistic

$$S_n := \frac{\sum_{t=1}^n R_t}{(\sum_{t=1}^n \hat{\sigma}^2(t))^{1/2}}$$

should be approximately standard normal. The p-value of the corresponding two-sided test is roughly equal to 1, and thus we accept the estimated distributions $\mathbf{G}_{\hat{\theta}(t)}$ as being

Test	Kolmogorov-Smirnov	Jarque-Bera	Shapiro-Wilk
p	0.897	0.336	0.265
p_{MC}	0.794	0.824	0.756

Table 4.1. *The p-values and Monte Carlo p-values for the KS, JB and SW tests for normality of the standardized returns.*

aproximately centered.

Our approach of evaluating the estimated model is not unproblematic, as it has two major shortcomings: when testing for independence of the returns, we assume the estimated distributions to be good approximations for the true return distributions. Whereas the goodness of fit tests used for checking normality of the standardized returns presume the returns to be independent. We adopt this approach, because (to the best of our knowledge) there is no reasonable method for testing independence and the goodness of fit *separately*. Consequently, the outcomes of the above evaluations should be interpreted carefully.

The second shortcoming of our approach concerns the p-values of the above goodness of fit tests. What we did above, was computing the estimates $\hat{\theta}(t)$ on the basis of the centered returns and then applying the tests for normality on the standardized returns. These depend on the estimates $\hat{\theta}(t)$ and thus, again, on the centered returns (cf. (4.3)). So, actually, we use the same data for estimating the model as we use for checking their goodness of fit. Hence, the accuracy of the p-values seems questionable. Due to the high dependency on only one single set of data, they may not be very accurate and there might be some need for adjusting them. Therefor let us remind how the p-values are calculated for the case of the KS, JB and SW tests: each of these tests is based on a test statistic, say T. It will reject the hypothesis of standard normal V_t's if the realization of T, denoted t, takes large values. The p-value is then defined as $p = P\{T \geq t\}$. Thus, we can redress the problem of an inaccurate p-value by recalculating the law of T in consideration of the standardization (4.3). To this end, suppose we are given J time series samples $\{R_t^{*j} : t = 1, \ldots, n\}$, simulated according to (4.2). For each of these simulated samples we compute the likelihood estimates $\{\hat{\theta}^{*j}(t) : t = 1, \ldots, n\}$ and the resulting standardized returns, given by $V_t^{*j} = \Phi^{-1}\big(\mathbf{G}(R_t^{*j}; \hat{\theta}^{*j}(t))\big)$. Consequently, we have $J + 1$ realizations of T given by

$$t = T(V_1, \ldots, V_n) \quad \text{and} \quad t_j^* = T(V_1^{*j}, \ldots, V_n^{*j}).$$

Then, from the empirical cdf $\mathbb{F}_J(\cdot) = J^{-1} \sum_{j=1}^{J} \mathbb{1}\{t_j^* \leq \cdot\}$ we get an approximation for the law of T under the assumption that the distributions of the centered returns are really given by the family $\{\mathbf{G}_{\hat{\theta}(t)} : t = 1, \ldots, n\}$. Consequently, for the p-value we have as an approximation the Monte Carlo p-value $p_{\mathrm{MC}} = 1 - \mathbb{F}_J(t)$. For more general information on this Monte Carlo testing approach cf. Davison and Hinkley (2003), especially Chapter 4. (Of course, for the same reasons that we already mentioned in the discussion of the construction of bootstrap confidence intervals, such a Monte Carlo p-value has to be considered carefully— first of all it is a heuristic approach to assess the quality of the model fit.)

We do this computationally intensive method for the KS, JB and SW tests with $J = 1000$ simulated times series. The results are given in the second line of Table 4.1. All three p_{MC}-values assert the good results from the "simple" tests, the Monte Carlo version of the JB and SW tests, surprisingly, having even a higher p-value than their simple versions.

Finally we want to check whether the approximate confidence intervals actually cover the true parameter with probability 0.95. Suppose that at time t neither $\hat{u}_+(t)$ nor $\hat{u}_-(t)$ equals 0 or 2/3, such that we can construct confidence intervals for the true parameters, according to Section 3. Again, we consider the simulated time series $\{R_t^{*j} : t = 1, \ldots, n\}$ and the resulting local likelihood estimates $\{\hat{\theta}^{*j}(t) : t = 1, \ldots, n\}$, $j = 1, \ldots, J$. Assume that from the estimates $\hat{\theta}^{*j}(t)$ (i.e. the parameter estimates for time t resulting from the jth
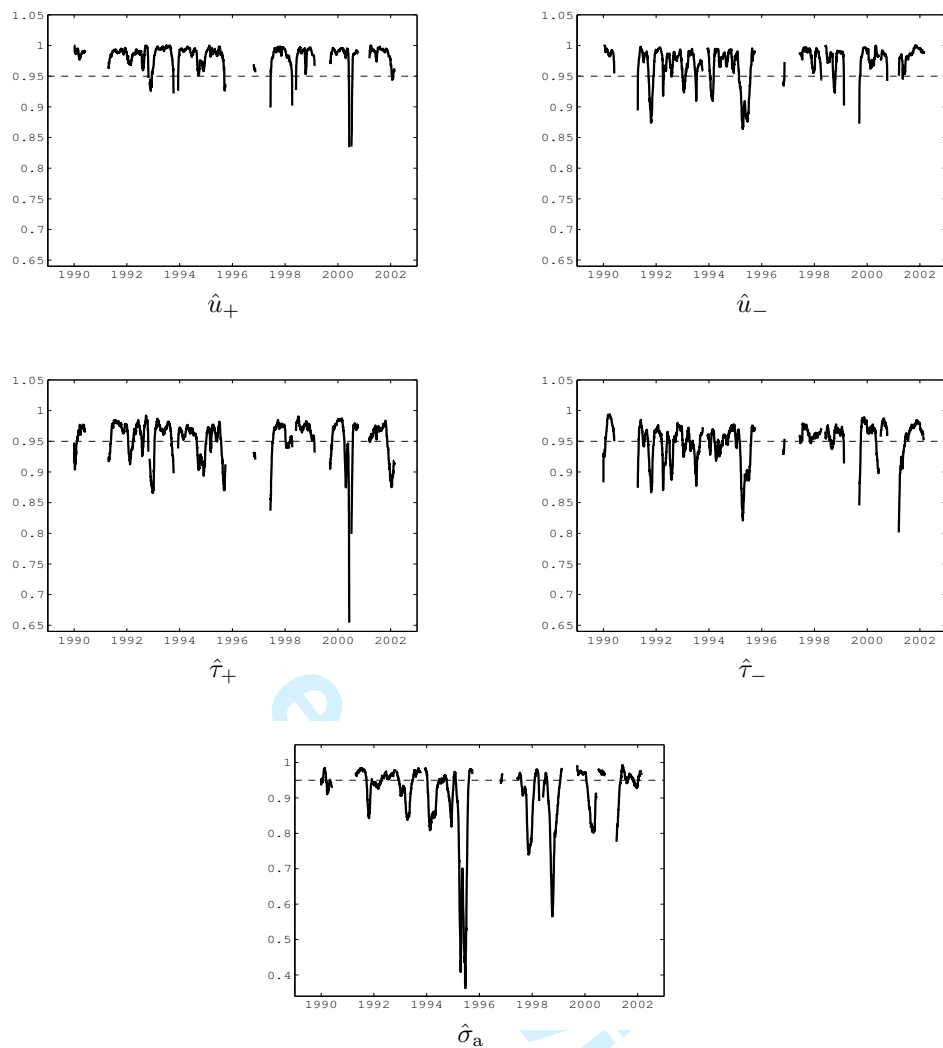
Figure 4.6. *Relative frequencies of the initial paramter estimates $\hat{\theta}_k(t)$, $k = 1, \ldots, 4$, to lie in the simulated confidence intervals. Points of time t, for which in the orignial estimates no confidence intervals are constructed, are assigned the value zero.*

simulation) we can compute approximate confidence intervals for each of the four parameters $\hat{\theta}_1(t), \ldots, \hat{\theta}_4(t)$. These may be denoted by $\mathcal{I}\big(\hat{\theta}_k^{*j}(t)\big)$, $k = 1, \ldots, 4$. The number of all j's for which the construction of such confidence intervals is possible be denoted by J_t. Then the relative frequencies $\#\big\{j : \hat{\theta}_k(t) \in \mathcal{I}\big(\hat{\theta}_k^{*j}(t)\big)\big\}/J_t$ give an estimate for the covering probability of the approximate confidence intervals.

We compute these relative frequencies for all t with $\hat{u}_+(t), \hat{u}_-(t) \notin \{0, 2/3\}$, using a total of $J = 1000$ simulated time series. The results are depicted in Figure 4.6: for the parameters u_+, u_-, τ_+ and τ_- the levels of the confidence intervals lie between 0.85 and 1.0 for almost all points of time and thus do not differ too much from the 95%-level. However, there are also some exceptions from this, e.g. shortly after the beginning of the year 2000, where the estimated level for the confidence intervals of the parameter τ_+ are very low (below 70%). A simple explanation for this effect might be, that this is exactly the period during which the estimated parameters $\hat{\tau}_+$ are maximal, as can be seen in Figure 4.1. The estimates $\hat{\tau}_+^{*j}$,

Test	Kolmogorov-Smirnov	Jarque-Bera	Shapiro-Wilk
p	0.790	0.812	0.861
p_{MC}	0.787	0.834	0.879

Table 4.2. *The p-values and Monte Carlo p-values for the KS, JB and SW tests for normality of the standardized returns when u_+ and u_- are modelled as global constants.*

which result from the simulated time series, in general take more moderate values, they are smaller than $\hat{\tau}_+$. Thus, the resulting simulated confidence intervals $\mathcal{I}(\hat{\tau}_+^{*j})$ cover the parameter $\hat{\tau}_+$ less often. For the annualized volatilities there are major differences between the relative frequencies and the presumed 95%-level of the confidence intervals. Especially the very low level shortly before 1996 is striking. This is exactly the period where there is the unexplainable double-peak of the volatilities (cf. Figure 4.1 and the comments thereon), and thus should not be too surprising. Note that, although we have taken $J = 1000$, of course the quantity J_t is less than that for all t, due to the fact that also during the simulations the normal approximation for the tails is used. However, except for very few points of time, we always have $J_t \geq 300$.

The model with globally constant parameters u_+ and u_-. In our first discussion of the estimates for the parameter functions u_+ and u_- (see p. 10), we pointed out that an alternative approach might be to model these two parameters by global constants. We want to briefly investigate how this simplified modelling approach behaves. To this end we take the estimated scale parameters $\hat{\tau}_+$ and $\hat{\tau}_-$ from the above local likelihood estimation and then we first re-estimate u_+ and u_- as global constants, using a maximum likelihood (ML) estimator. Based on these ML estimates, then the parameter functions τ_+ and τ_- are re-estimated with local likelihood estimates, using again local linear fits. Considering the S&P 500 data this yields globally constant parameter estimates $\hat{u}_+^{\mathrm{glob}} = 0.209$ and $\hat{u}_-^{\mathrm{glob}} = 0.315$, the newly estimated parameter functions $\hat{\tau}_+$ and $\hat{\tau}_-$ resemble very much the corresponding functions from Figure 4.1, the plots are therefore omitted. The same model assessment procedures that are used above can now be applied to the standardized returns resulting from these new parameter estimates: the assessments of the SACFs and the Portmanteau tests look exactly the same as those from Figures 4.3 and 4.5 (and are therefore omitted), thus indicating that the parameter estimates describe the data quite well. This is further supported by the p-values and the Monte Carlo p-values—based on $J = 1000$ simulated time series—of the goodness of fit tests, which are given in Table 4.2. The p-values from the simplified modelling approach even seem to slightly outperform the results from Table 4.1.

Of course, the above "mixed" modelling approach, where global constant fits for u_+ and u_- and local linear fits for τ_+ and τ_- are used, comes with the drawback that nearly two times as many parameter estimates are needed in order to fit the model. However, it has the advantage that—if once fitted—the model can be updated more easily, because only the scale parameters τ_+ and τ_- have to be updated regularly, which thus also saves on computing time. This is in particular interesting for real-time observation of a time series.

5 Forecasting with local likelihood methods

We now want to provide an approach for forecasting future return distributions, the basic idea of which is the same as in Drees and Stărică (2002). Subsequently, we provide numerical results for one-day-ahead distributional forecasts, and discuss them.

Distributional forecasts in the regression model. Assume we are given the data available up to time t—which we interprete as the present—and we try to forecast the distribution of the next d-day return $X_{t,d} = \sum_{i=1}^d X_{t+i}$. In our model this is given by

$$
(5.1) \qquad
\begin{cases}
X_{t,d} = d\mu + \sum_{s=1}^d R_{t+s}, \\[2mm]
R_{t+s} \sim \mathbf{G}_{\theta(t+s)}, \quad R_{t+s} \text{ independent}, \; s = 1, \dots, d.
\end{cases}
$$

Thus, we have to find reasonable estimates for the future parameter values $\theta(t+s)$ that are based only on the data available at time t. Presuming the forecasting horizon d is not too large, an extrapolation of the Taylor polynomial in (3.1)—using the notation from Section 3, especially the vector notation from (3.2)—yields

$$
(5.2) \qquad \theta_k(t+s) \approx \sum_{j=0}^{p_k} \frac{\theta_k^{(j)}(t)}{j!} s^j \equiv \sum_{j=0}^{p_k} \beta_{kj}(t) s^j, \quad s = 1, \dots, d, \quad k = 1, \dots, 4.
$$

The unknown coefficients $\beta_{kj}(t)$ can be estimated by local likelihood methods. Therefor we employ a one-sided kernel, i.e. we use $K(\cdot)\mathbb{1}\{\cdot \le 0\}$ instead of $K(\cdot)$. Again, the estimations are based on the centered returns, which in the forecasting context are given by $\tilde{R}_s = X_s - \overline{X}_t$, $s \le t$, where $\overline{X}_t = 1/t \sum_{i=1}^t X_i$. (Note that these centered returns have to be re-computed for different values of t.) These two modifications assure that the resulting local likelihood estimates $\hat{\beta}_{kj}(t)$ only depend on the information available at time t. Plugging these initial estimates into (5.2), we get forecasts for the future parameters by

$$
(5.3) \qquad \hat{\theta}_{k,t}(t+s) = \sum_{j=0}^{p_k} \hat{\beta}_{kj}(t) s^j, \quad s = 1, \dots, d, \quad k = 1, \dots, 4.
$$

Here and in the following the additional subscript t shall emphasize that the parameter forecast is only based on the data up to time t. Instead of (5.3), one might also take

$$
(5.4) \qquad \hat{\theta}_{k,t}(t+s) = \hat{\beta}_{k0}(t), \quad s = 1, \dots, d, \quad k = 1, \dots, 4,
$$

thus neglecting any trend components, which express through the components $\hat{\beta}_{kj}(t)$, $j > 0$. In the case of fitting each parameter by local constants, i.e. $p_1 = \dots = p_4 = 0$, then of course (5.3) and (5.4) coincide. Replacing the unknown quantities μ and $\theta(t+s)$ in (5.1) by $\overline{X}_t$ and $\hat{\theta}_t(t+s)$ (either given by (5.3) or by (5.4)), the law of

$$
(5.5) \qquad \tilde{X}_{t,d} = d\overline{X}_t + \sum_{s=1}^d \tilde{R}_{t+s}, \quad \tilde{R}_{t+s} \sim \mathbf{G}_{\hat{\theta}_t(t+s)},
$$

provides a distributional forecast for $X_{t,d}$. In the following, the correspondig cdf's of $X_{t,d}$ and $\tilde{X}_{t,d}$ are denoted $F_{t,d}$ and $\tilde{F}_{t,d}$, respectively.

The asymptotic results for the local likelihood estimates presented in Section 3 act on the assumption of a symmetric kernel, and hence seem not suitable for our forecasting purpose.

However, they can easily be carried over to our forecasting context. To this end, for each fixed t we interpret t as the right boundary point of the interval $[1, t]$ and we do the rescaling proposed in (3.4) with n replaced by t. The bandwidth is rescaled in a similiar way, taking $\tilde{h} = h \cdot n/t$ for the estimation of $\theta(t)$, with h fixed. The rescalation of the bandwidth ensures that the number of observations used for the estimation does not change with varying t. With this interpretation of t as right boundary point, it makes no difference if we use the above one-sided kernel or its two-sided version, because there are no observations at times larger than t that could have influence on the estimate. Consequently, the asymptotics, which hold true for the right boundary point t when using a two-sided kernel, do so for the use of the one-sided kernel as well. That is, the estimates $\hat{\beta}(t)$ are consistent and asymptotically normal, which enables us to provide approximate confidence intervals for the true parameters (as proposed in Section 3). Again, it is favourable to use local polynomials of the same degree for all parameters, i.e. $p_1 = \cdots = p_4$, in order to keep the expression for the asymptotic bias to a minimum. For the fitting of the model it was also preferable to chose odd degree local polynomials to ensure that the expression for the bias is of the same order for boundary points and inner points. Since in this situation we always consider the estimation of the parameter function in a boundary point, the question whether the p_i's should be even or odd is of minor importance (cf. Aerts and Claeskens, 1997 for more details).

We apply the above forecasting approach to the S&P 500 daily returns, and we do one-day-ahead distributional forecasts, i.e. we take $d = 1$. Two different computations are done, each using a different degree of local polynomial fitting, namely:

(a) All four parameters are fitted by local constants, i.e. $p_1 = \cdots = p_4 = 0$.

(b) Local linear fits are used for each of the four parameters, i.e. $p_1 = \cdots = p_4 = 1$.

For each of the two fitting schemes (a) and (b) we implement the cross-validation method (cf. Section 3), computing the corresponding cross-validation functions for bandwidths ranging from 0.040 to 0.200. In case of fitting scheme (a) the resulting function is close to constant, with a maximum at $h_{\mathrm{CV}} \approx 0.13$, which is however not very distinct. For fitting scheme (b) the corresponding cross-validation function is strictly increasing for all considered bandwidths. On the basis of these results it seems hard to choose an appropriate bandwidth, in both cases the cross-validation method yields large bandwidths. Indeed, it is plausible that for the prediction of future parameters a larger degree of smoothing is needed than for the fitting of the model. However, in case of fitting scheme (b) this is somewhat critical: the bigger we choose h the more restrictive becomes the second constraint in (4.1), and thus the smaller in absolute value the estimates $\hat{\theta}_k^{(1)}$ for the first derivatives of the parameter functions have to be, which runs contrary to the idea of local linear fits for the parameter functions. Therefore, in this situation it seems questionable if cross-validation is the best approach for the choice of an appropriate bandwidth. Instead, we choose a smaller bandwidth for our computations, taking $h = 0.095$, which is then rescaled for each t, as described above. This bandwidth choice is essentially based on trial and error, doing parameter estimations with different bandwidths.

We use the one-sided version of the Epanechnikov kernel. Otherwise, the same setting—i.e. same parameter space, optimization procedure, etc.—as in the simple model fitting

$\hat{u}_+$

$\hat{u}_-$

$\hat{\tau}_+$

$\hat{\tau}_-$

$\hat{\sigma}_{\mathrm{a}}$
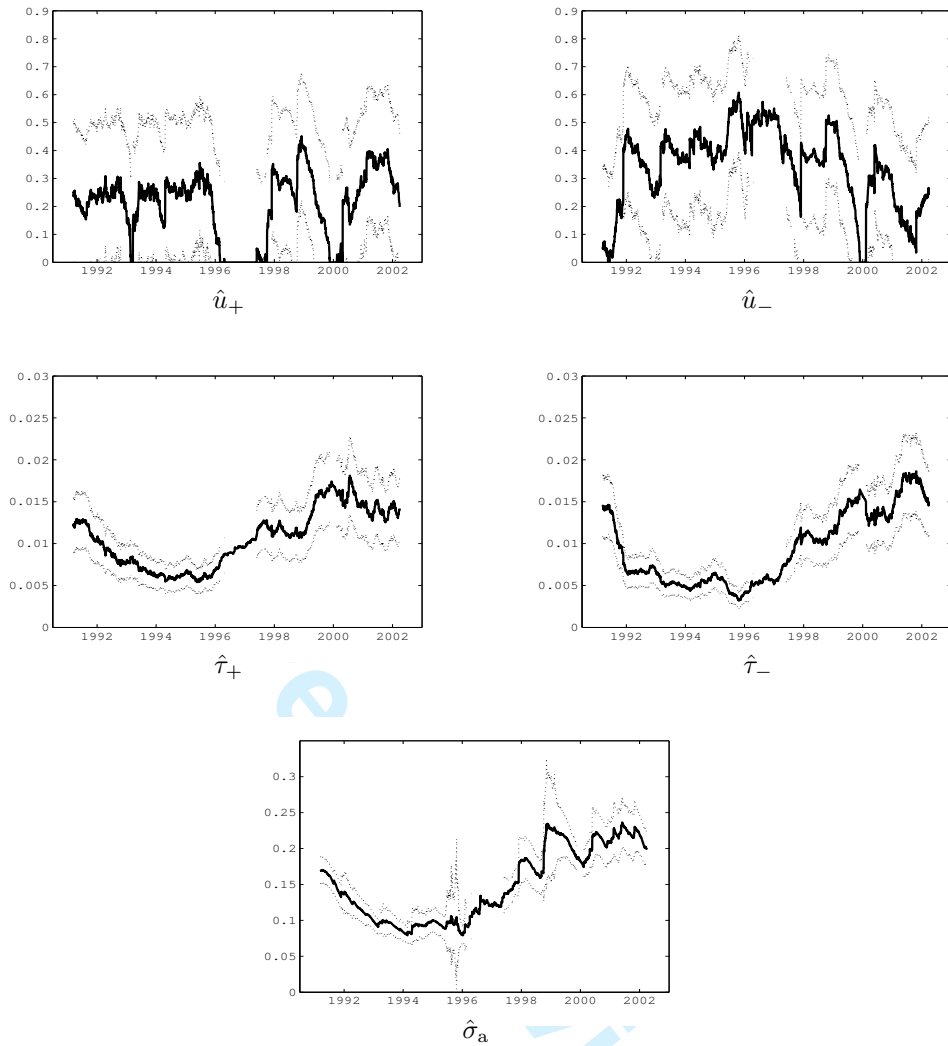
Figure 5.1. *The estimated parameter functions $\hat{u}_+$, $\hat{u}_-$, $\hat{\tau}_+$ and $\hat{\tau}_-$, and the resulting annualized volatilities $\hat{\sigma}_{\mathrm{a}}$ for the case of local constant fits (case (a)). The estimated functions are depicted in solid lines, the corresponding approximate pointwise 95%-confidence intervals (where applicable) are depicted in dotted lines.*

approach is used (cf. Section 4). Due to the use of a one-sided kernel and the specific choice of bandwidth we do not compute local likelihood estimates for $t \leq 300$.

Numerical results for the forecasts. Figure 5.1 shows the local likelihood estimates $\hat{\theta}_{k,t}(t)$ for $t = 301, \ldots, 3062$, and the corresponding 95%-confidence intervals (where applicable) for the case of local constant fits of all four parameters (case (a)). The estimated functions, and thus the estimated distributions, seem to exhibit the same symmetric behaviour that was already discussed in Section 4. Furthermore, for almost all points of time, the left tails—i.e. the tails modelling the losses—are heavier than the right tails, a desirable property, as mentioned above. Despite the much larger bandwidth, the estimated functions in Figure 5.1 exhibit greater fluctuations on a small time scale than the corresponding estimates for the fitting of the model (Figure 4.1). This is due to the use of the one-sided kernel.

We omit plots of the estimated parameter functions for case (b): in this case the estimated functions look similar to those from Figure 4.1, however they are much more wiggly, and the confidence intervals are wider.

For the prediction of the law of the one-day-ahead returns $X_{t,1}$ we consider three different forecasting approaches. These differ with respect to the degrees of the local polynomials that are used for the initial likelihood estimations (represented by one of the above cases (a) and (b)), and with respect to the kind of extrapolation that is used for predicting future distribution parameters (either (5.3) or (5.4)):

F1: The initial local likelihood estimates are computed with local constant fits for all parameters. In this case (5.3) and (5.4) are the same.

F2: For the initial likelihood estimates local linear fits are used for all parameters, the parameter forecasts are given by (5.3).

F3: Same as F2, but with parameter forecasts according to (5.4).

We now want to analyse the quality of the distributional forecasts provided by the random variables $\tilde{X}_{t,1}$. Essentially, the analysis is based on the same ideas and testing devices that we used for the assessments of the quality of the model fitting in Section 4. Suppose for a moment that the forecasting distributions $\mathscr{L}(\tilde{X}_{t,1})$ perfectly match the laws of the true future returns $\mathscr{L}(X_{t,1})$, i.e. for the corresponding cdf's we have $\tilde{F}_{t,1}(x) = F_{t,1}(x)$ for all $x \in \mathbb{R}$. Then the variables

$$(5.6) \qquad Z_{t,1} := \Phi^{-1}\big(\tilde{F}_{t,1}(X_{t,1})\big), \quad t = 301,\ldots,3061,$$

—in the following referred to as *standardized forecasts*—should be i.i.d. standard normal. To check whether this holds true, we have to test for both, independence and normality of the standardized forecasts. We do so for each of the forecasting methods F1–F3.

In a first step to check for the normality of the standardized forecasts, we apply the goodness of fit tests (KS, JB ans SW tests) to the time series $\{Z_{t,1} : t = 301,\ldots,3061\}$. The corresponding p-values for each of the three forecasting approaches are given in Table 5.1. Considering only the p-values resulting from the simple versions of the three goodness of fit tests does not allow for an answer to the question whether the standardized forecasts are (at least approximately) standard normal: as can be seen from the table, the KS test supports the hypothesis of normality of the standardized forecasts for all three forecasting approaches, whereas the JB and SW tests nearly always reject the hypothesis. This conflict might be a consequence of the standardization (5.6) and its effects on the values of the different test statistics (compare the discussion in Section 4), thus indicating the inadequacy of these simple goodness of fit tests. For this reason, again, we consider Monte Carlo tests. The basic idea is the same as described in Section 4: we simulate J time series $\{\tilde{R}_t^{*j} : t = 301,\ldots,3062\}$, with independently simulated random variables $\tilde{R}_t^{*j} \sim \mathbf{G}_{\hat{\theta}_t(t)}$. Based on these simulated time series we re-compute the initial local likelihood estimates $\{\hat{\beta}^{*j}(t) : t = 601,\ldots,3062\}$, and we compute the corresponding forecasting parameters $\hat{\theta}_t^{*j}(t+1)$ and the simulated standardized forecasts $\{Z_{t,1}^{*j} : t = 601,\ldots,3061\}$. Repeating these operations for $j = 1,\ldots,J$, and computing the KS, JB and SW test statistics for these simulated standardized forecasts, finally yields approximate Monte Carlo p-values.

| Forecasting method	Test	Kolmogorov-Smirnov	Jarque-Bera	Shapiro-Wilk
F1	p	0.969	0.003	0.054
	p_{MC}	0.902	0.656	0.750
F2	p	0.235	0.000	0.000
	p_{MC}	0.906	0.703	0.744
F3	p	0.244	0.000	0.000
	p_{MC}	0.902	0.701	0.741

Table 5.1. *The p-values and Monte Carlo p-values for the KS, JB and SW tests for the standardized forecasts resulting from the three forecasting methods F1–F3. The Monte Carlo p-values are based on 1000 simulated time series.*
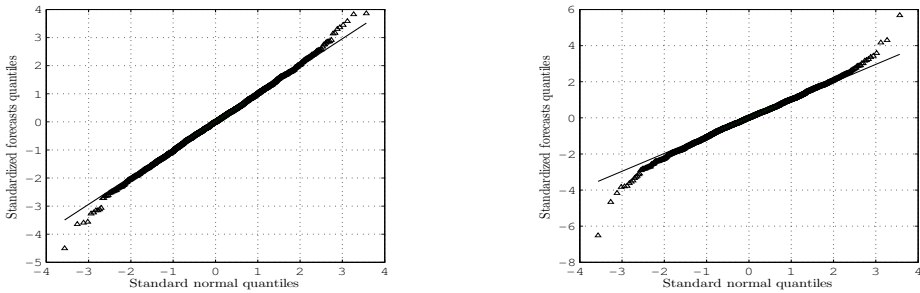
Figure 5.2. *QQ-plots for the standardized forecasts $\{Z_{t,1} : t = 301, \ldots, 3061\}$ resulting from the forecasting approaches F1 (left) and F3 (right).*

We do these Monte Carlo tests for each of the forecasting approaches F1–F3, where each of the computed p_{MC}-values is based on $J = 1000$ simulated time series. The results are given in Table 5.1, too. Without exception, the tests strongly support the hypothesis of normality of the standardized forecasts, and thus the idea of a good forecasting. Apparently, there are only small differences in the p_{MC}-values for Methods F2 and F3. Thus, for the quality of one-day-ahead parameter forecasts it seems not of great importance whether these incorporate a trend component or not. This is due to the fact that in our special case the values of the forecasting parameters $\hat{\theta}_{k,t}(t+1)$ only vary at an order of magnitude of 10^{-3}, depending on whether they are computed according to (5.3) or (5.4). Consequently, for the one-day-ahead distributional forecasts it has nearly no effects which of these two methods is used for the computation of $\hat{\theta}_{k,t}(t+1)$. For the same reason one can anticipate that the forecasting results will not improve significantly by chosing higher order polynomial fits, and therefore—for one-day-ahead forecasts—the use of local constant fits seems favourable since it is computationally least intensive.

Next we consider the QQ-plots of the standardized forecasts. For the cases F1 and F3 these are depicted in Figure 5.2. (The QQ-plot for the standardized forecasts based on forecasting method F2 looks very similar to that of F3, so we omit it.) As can be seen from the figure, the data roughly fit the main diagonal, thus underpinning the normality hypothesis for the standardized forecasts. Solely in the tails there seem to be some deviations, especially when considering the standardized forecasts resulting from method F3 (Figure 5.2, right).
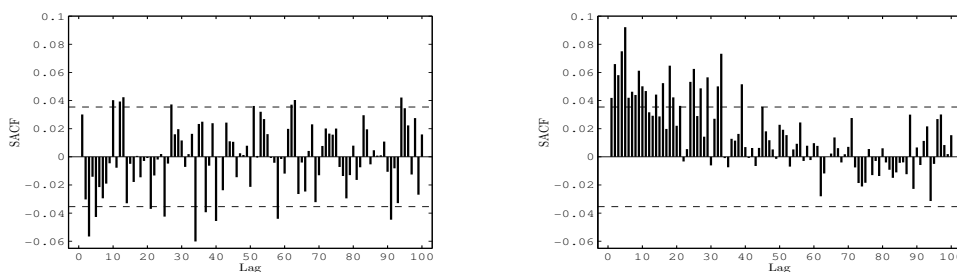
Figure 5.3. *The SACFs of the standardized forecasts $\{Z_{t,1} : t = 301, \ldots, 3061\}$ (left) and of the absolute values $\{|Z_{t,1}| : t = 301, \ldots, 3061\}$ (right), each for the first 100 lags. The standardized forecasts are based on the forecasting approach F1. The dashed lines represent the 95%-confidence intervals.*

Altogether, taking into account both the QQ-plots and the results of the Monte Carlo goodness of fit tests, the data seem not to directly contradict the hypothesis of normality of the standardized forecasts—regardless of which forecasting approach F1, F2 or F3 is considered. Thus, at least approximately the standardized forecasts can be considered normal. This supports the assumption that the distributional one-day-ahead forecasts $\tilde{F}_{t,1}$ are reasonable approximations for the true distributions $F_{t,1}$.

It remains to investigate the dependence structure of the standardized forecasts. To this end, we consider the SACFs of the time series $\{Z_{t,1}\}$ and $\{|Z_{t,1}|\}$. For the forecasting approach F1 these are depicted examplarily in Figure 5.3. The corresponding SACFs for the standardized forecasts based on the other forecasting approaches look very similar to the ones showed in Figure 5.3 and thus are omitted. Under the assumption of independence of the standardized forecasts, the absolute values of both SACFs should be very small for all lags. The SACF of the standardized forecasts $\{Z_{t,1}\}$ clearly exhibits such a behaviour and thus might be evidence for the independence assumption to be true. However, the SACF of the time series $\{|Z_{t,1}|\}$ clearly seems to differ from that. For about the first 30 lags the SACF exceeds the upper 95%-confidence bound. In addition, the Portmanteau tests for both, the standardized forecasts as well as for their absolute values, reject the hypothesis of independence for almost all lags. The corresponding p-value plots look exactly the same as the top right image in Figure 4.3, and thus are omitted. (The statements just made also hold true for the other forecasting approaches F2 and F3.) Thus, the hypothesis of the standardized forecasts to be independent seems a bit questionable.

Forecasting over longer horizons. Of course the forecasting approach that we used above can also be used for larger forecasting horizons, say $d = 20$ or $d = 40$. Drees and Stărică (2002) do such examinations. However, with a larger forecasting horizon there occur several difficulties: one major problem is that the distributions of the forecasting variables $\tilde{X}_{t,d}$ cannot be described in a closed form. Hence, simulations are needed to find good approximations for the convolution given in (5.5). Furthermore, the variables $X_{t,d}$ are not independent. Even in case of the returns being independent—which is a somewhat critical assumption, as we have seen in the above examinations—the forecasting variables

form a $(d-1)$-dependent sequence of random variables. And the same holds true for the corresponding standardized d-day forecasts $Z_{t,d} = \Phi^{-1}\big(\tilde{F}_{t,d}(X_{t,d})\big)$. Hence, the basis for an application of the goodness of fit tests is clearly violated. (Remember that this was already critical for the one-day-ahead forecasts!) On the other hand Monte Carlo tests, like we did them above, are computationally way too expensive, because of the additional simulations needed for computing the distributions of the variables $X_{t,d}$. Hence, an assessment of the forecasts becomes a difficult task. For these reasons we only consider one-day-ahead forecasts in this paper.

6   Concluding remarks

Based on a work of Drees and Stărică (2002), we considered a simple regression type model for stock returns. The innovations were modelled by four parameter asymmetric distributions, and the parameters were modelled as smooth, deterministic functions of time. In addition, we considered an appropriate normal approximation for the tails, enabling the model to adapt to both, heavy-tailed as well as light-tailed return distributions. By the example of the S&P 500 daily returns, observed over a period of twelve years, the principle of local likelihood estimation (in connection with local linear fits) prove to be a reasonable method for fitting our model to real-life data. In addition, approximate confidence intervals for the true parameters could be given. The assessment of the model via the (Monte Carlo) KS, JB and SW tests argued for a good fitting of our estimated model to the given data. Solely the assumption of independence of the innovations seemed somewhat critical: although the investigation of the SACFs of the (standardized) returns and their absolute values supported the hypothesis of independent innovations, the outcome of the Portmanteau tests partly spoke against that. We also considered a simple device for forecasting future return distributions, and we did one-day-ahead distributional forecasts, using local likelihood estimators with several forecasting approaches. Actually, an evaluation of these estimates pointed out some minor flaws, concerning the independence assumption of the returns and the results of the simple versions of the goodness of fit tests. However, the results of the Monte Carlo tests as well as the QQ-plots for the standardized forecasts clearly argued for our modelling approach to yield reasonable distributional forecasts.

Summarizing, one can say that our simple modelling approach, in connection with the local likelihood estimators, seems to succeed in describing financial time series and giving reasonable one-day-ahead forecasts. The results clearly underpin the adequacy of the idea of a local modelling approach.

References

Aerts, M. and Claeskens, G. (1997). Local polynomial estimation in multiparameter likelihood models. *Journal of the American Statistical Association*, 92:1536–1545.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327.

Bollerslev, T., Engle, R., and Nelson, D. (1994). ARCH models. In Engle, R. and McFadden, D., editors, *Handbook of Econometrics, vol. 4*, pages 2961–3038. Elsevier.

Brockwell, P. and Davis, R. (1987). *Time Series Analysis: Theory and Methods*. Springer, New York.

Davison, A. and Hinkley, D. (2003). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Drees, H. and Stărică, C. (2002). A simple non-stationary model for stock returns. Preprint, available at www.math.chalmers.se/~starica.

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, 50:987–1008.

Fan, J., Farmen, M., and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society, series B*, 60(3):591–608.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall, London.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series*. Springer, New York.

Gouriéroux, G. (1997). *ARCH Models and Financial Applications*. Springer, New York.

Jönck, U. (2005). Lokale Likelihood-verfahren zur Modellierung von Finanzzeitreihen. Diplomarbeit, Universität Hamburg.

Johnson, N. and Kotz, S. (1970). *Continuous Univariate Distributions, vol. 2*. Houghton Mifflin, Boston.

Judge, G., Hill, R., Griffiths, W., Lütkepohl, H., and Lee, T.-C. (1988). *Introduction tho the Theory and Practice of Econometrics*. Wiley, New York.

Mikosch, T. and Stărică, C. (2003). Long-range dependence effects and ARCH modelling. In Doukhan, P., Oppenheim, G., and Taqqu, M., editors, *Theory and Applications of Long-range Dependence*, pages 439–459. Birkhäuser, Boston.

Mikosch, T. and Stărică, C. (2004). Non-stationarities in financial time series, the long range dependence and the IGARCH effect. *Review of Economics and Statistics*, 86:378–390.

Shapiro, S. and Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611.

Stărică, C. (2003). Is GARCH(1,1) as good a model as the Nobel prize accolades would imply? Preprint, available at www.math.chalmers.se/~starica.