Smoothing long memory time series using R

Yuanhua Feng, Jan Beran and Sebastian Letmathe
Faculty of Business Administration and Economics, Paderborn University

May 29, 2021

Abstract

The paper at hand provides a detailed description of the *smootslm* R-package, which is an extension of the already published *smoots* package, enabling the data-driven local-polynomial smoothing of time series with long-memory. In this regard a simple data-driven algorithm is proposed based on the well-known iterative plug in algorithm for SEMIFAR (semiparametric fractional autoregressive) models. Two new functions for data-driven estimation of the trend and its derivatives under the presence of long-memory are introduced. *smootslm* is applied to various environmental and financial time series with long memory, e.g. mean monthly Northern Hemisphere changes and annual GER-GDP. Moreover, *smootslm* is capable of estimating semiparametric FI-Log-GARCH and FI-Log-ACD models. It is worth mentioning that this package can be applied to any suitable time series with long memory.

Keywords: long-memory, data-driven smoothing, SEMIFAR, estimation of derivatives, *smootslm* package

JEL Codes: C14, C51

1 Introduction

This paper introduces a new R-package coined *smootslm* designed to supplement the already published *smoots* package (smoothing time series, version 1.0.1, Feng and Schulz, 2019). The latter considers data-driven local polynomial smoothing of trend-stationary time series with short range dependence structure. However, the literature suggests that time series, for instance squared returns or trade durations, clearly exhibit long memory (see e.g. Ding et al., 1993, Ding and Granger, 1996, Andersen and Bollerslev, 1997, Andersen et al., 1999, Cotter, 2005 and Beran et al., 2015 among others).

Against this background the *smootslm* package is developed enabling the data-driven trend estimation under long memory. Analogously to *smoots* the estimation of the trend and its first and second derivative is carried out by means of a data-driven IPI (iterative plug-in Gasser et al., 1991) method based on the IPI for SEMIFAR (semiparametric fractional autoregressive, Beran and Feng, 2002c) models introduced by Beran and Feng (2002a). The SEMIFAR and its exponential version the ESEMIFAR model (Beran et al., 2015), which is applicable to non-negative time series following a semiparametric multiplicative model form, are designed for simultaneous modelling of stochastic trends, deterministic trends and stationary short- and long-memory components in a time series.

The theoretical background of the *smootslm* package and its implementation in R are briefly exemplified. For further details on the theoretical properties of the (E)SEMIFAR model and the corresponding IPI-algorithm we refer the reader to Beran and Ocker (1999), Beran and Feng (2002c), Beran and Feng (2002a), Beran and Feng (2002b), Beran et al. (2015), Beran et al. (2016) and references therein. The main objective of this paper is the introduction of the *smootslm* package and the illustration of its usefulness in particular for non-stationary time series exhibiting a long-memory dependence structure. That is achieved by employing our package to different environmental as well as financial time series. We partly exploit data that was already used by Feng et al. (forthcoming), namely monthly Northern Hemisphere temperature changes. Analogously to Feng et al. (forthcoming) *smootslm* is used in the context of a semiparametric log-local-linear growth model for analysing annual german GDP data. Furthermore, our proposal is applied within the scopes of the Semi-FI-Log-GARCH model which was recently introduced by Feng et al. (forthcoming). The authors added a nonparametric scale function into the FI-

Log-GARCH proposed by Feng et al. (2020), which is in turn a long memory extension of the Log-GARCH (Pantula (1986), Geweke, 1986 and Milhøj, 1987) and is closely related to conventional long memory GARCH models. The former is applied to a return series of the S&P500. Moreover, it was first indicated by Beran et al. (2015) that the (type 1) Log-ACD model introduced by Bauwens and Giot (2000), Bauwens et al. (2008) and Karanasos (2008) can be represented as an EFARIMA model. Subsequently, it was shown by Feng and Zhou (2015) that the EFARIMA and ESEMIFAR can be redefined as a FI-Log-ACD and a Semi-FI-Log-ACD, respectively. We illustrate the use of *smootslm* in regards to the Semi-FI-Log-ACD by modelling log-transformed trading volume of the S&P500.

The paper is organised as follows. In Section 2 the definitions of the FARIMA and SEMIFARIMA are given and the methodological background as well as the IPI-algorithm incorporated in *smootslm* are elaborated. The implementation in R is exemplified in Section 3. The application of our proposal to environmental data is illustrated in Section 4. The use of *smootslm* within the scopes of the Semi-FI-Log-GARCH and Semi-FI-Log-ACD is investigated in Sections 5 and 6. In Section 7 final remarks are given.

2 Smoothing long memory time series

2.1 The FARIMA and SEMIFARIMA

A well-established model for analysing financial time series data is the multiplicative error model (MEM) (Engle, 2002) which is given by

$$X_t = s\lambda_t\eta_t, \tag{1}$$

where the scale parameter is denoted by $s > 0$, $\lambda_t > 0$ denotes the conditional mean of $X^* = X_t/s$, and η_t are i.i.d. random variables with zero mean and unit variance. Following Feng and Zhou (2015) we can rewrite (1) as a semiparametric MEM given by

$$X_t = s(\tau_t)\lambda_t\eta_t, \tag{2}$$

where $\tau_t = t/n$ denotes the rescaled time and where the scale parameter s in (1) is replaced with a nonparametric scale function denoted by $s(\tau_t)$. By taking the logs of (2) we have

$$Y_t = g(\tau_t) + Z_t, \tag{3}$$

where $Y_t = \ln(X_t)$, $g(\tau_t) = \ln[s(\tau_t)]$, $Z_t = \ln(\lambda_t) + \epsilon_t$ and $\epsilon_t = \ln(\eta_t)$. Following Beran and Feng (2002c) we assume that Z_t follows a zero mean FARIMA (p, d, q) process, which is given by

$$(1 - B)^d \phi(B) Z_t = \psi(B)\epsilon_t, \tag{4}$$

where $d \in (0, 0.5)$ is the long-memory parameter, B is the backshift operator, $\phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i$ and $\psi(z) = 1 + \sum_{i=1}^{q} \psi_i z^i$ are AR- and MA-polynomials with all roots outside the unit circle. Equation (4) defines a stationary and invertible FARIMA process with $E(\epsilon_t) = 0$ and $var(\epsilon_t) = \sigma_\epsilon^2$. Model (3) is equivalent to a SEMIFAR process (Beran and Feng, 2002c) with no integer differencing ($m = 0$) and an additional MA-part. Please note that $X_t^* = \exp(Z_t)$. Subsequently, model (2) is an extended version of an ESEMIFAR introduced by Beran et al. (2015). However, the authors assumed that X_t^* is log-normally distributed whereas in this paper we relax this assumption and suppose that X_t^* satisfies condition **A1** of Feng et al. (2020).

2.2 Local polynomial regression for long memory time series

In the following local polynomial estimation of the scale function $g^{(\nu)}$, the ν-th derivative of g, is exemplified briefly (see e.g. Beran and Feng, 2002a, Beran and Feng, 2002b, Beran and Feng, 2002c, and Beran et al., 2013). Under the assumption that g is at least $(l+1)$-times differentiable at a point t_0, $g(\tau_t)$ can be approximated by a local polynomial of order l for τ_t in a neighbourhood of τ_0. Following Gasser and Müller (1979), the weight function is determined to be a second order kernel with compact support $[-1, 1]$ having the polynomial form $K(u) = \sum_{i=0}^{r} a_i u^{2i}$, for ($|u| \leq 1$),where $K(u) = 0$ if $|u| > 1$ and a_i are such that $\int_{-1}^{1} K(u)du = 1$ holds. Here, $r \in \{0, 1, 2, 3\}$ denotes the kernel used for estimating $g^{(\nu)}$, corresponding to the uniform, epanechnikov, bisquare and triweight kernel, respectively. $\hat{g}^{(\nu)}$ ($\nu \leq l$) can now be obtained by solving the locally weighted least

squares problem

$$Q = \sum_{i=1}^{t} \left[Y_t - \sum_{j=0}^{l} b_j (\tau_i - \tau_0)^j \right]^2 K\Big(\frac{\tau_i - \tau_0}{h}\Big), \tag{5}$$

where h denotes the bandwidth and $K[(\tau_i - \tau_0)/h]$ are the weights ensuring that only observations in the neighbourhood of τ_0 are used. Consider the case where $l - \nu$ is odd. Define $m = l + 1$, then we have $m \geq \nu + 2$ and $m - \nu$ is even. A point τ is said to be in the interior for each $\tau_t \in [h, 1-h]$, at the left boundary if $\tau_t \in [0, h)]$ and at the right boundary if $\tau_t \in (1-h, 1]$. Following Beran and Feng (2002b) a common definition for an interior point is $\tau = ch$ with $c = 1$ and for a boundary point we have $c \in [0, 1)$. Asymptotic expressions for the bias, variance and mean integrated squared error (MISE) are presented in Theorem 1 and 2 by Beran and Feng (2002b). The asymptotic mean integrated squared error (AMISE) is given by

$$\text{AMISE}(h) = h^{2(m-\nu)} \frac{I[g^{(m)}]\beta^2}{m!} + \frac{(nh)^{2d-1} V(1)}{h^{2\nu}}, \tag{6}$$

where $I[g^{(m)}] = \int_{c_b}^{1-d_b} [g^m(\tau)]^2 d\tau$ with $0 \leq c_b < d_b \leq 1$ in order to reduce the so-called boundary effect. Moreover, $\beta = \int_{-1}^{1} u^m K(u) du$ and for $d > 0$ we have $V(1) = 2c_f \Gamma(1-2d) \sin(\pi d) \int_{-1}^{1} \int_{-1}^{1} K(x)K(y)|x-y|^{2d-1} dxdy$. For $d = 0$, V reduces to $V(1) = 2\pi c_f \int_{-1}^{1} K^2(x) dx$. c_f stands for the spectral density of the ARMA part of (4) at frequency zero and is given by

$$c_f = f(0) = \frac{\sigma_\epsilon^2}{2\pi} \frac{(1 + \psi_1 + \cdots + \psi_q)^2}{(1 - \phi_1 - \cdots - \phi_p)^2}. \tag{7}$$

The asymptotically optimal bandwidth, denoted by h_A, that minimizes the AMISE is given by

$$h_A = C n^{(2d-1)/(2m+1-2d)}, \tag{8}$$

with

$$C = \left(\frac{[m!]^2}{2(m-\nu)} \frac{(2\nu + 1 - 2d)}{\beta^2} \frac{(d_b - c_b) V(1)}{I[g^{(m)}]} \right)^{1/(2m+1-2d)}. \tag{9}$$

Based on these results Beran and Feng (2002a) proposed two iterative plug-in algorithms for automatic bandwidth selection, namely Algorithm **A** and **B**. In this paper we only consider a strongly adapted version of Algorithm **B** which is presented in the following.

In order to obtain a selected bandwidth the unknown constants $I[g^{(m)}]$, d and V in (8) have to be replaced with consistent estimators. Please note, that the estimation of

V relates to that of cf. $I[g^{(m)}]$ is estimated by means of local polynomial regression and numerical integration. The remaining two quantities d and V can be obtained via maximum likelihood. Inserting those estimates into (8) yields a plug-in estimator for the bandwidth, which minimises the MISE.

2.3 The IPI-algorithm for estimating g

We introduce an IPI-procedure for SEMIFARIMA models by translating and adapting the main features of the IPI for SEMIFAR models introduced by Beran and Feng (2002a) from the programming language S to R. The algorithm processes as follows:

- **i)** In the first iteration start with an initial bandwidth h_0 set beforehand and select p and q denoting the AR- and MA-order, respectively.
- **ii)** Estimate g from Y_t employing h_{j-1} and calculate the residuals $\tilde{Z}_t = Y_t - \hat{g}(\tau_t)$. Estimate d and V by fitting a FARIMA (with predefined AR- and MA-order in Step i) to $\hat{Z}_t$.
- **iii)** Set $h_{d,j} = (h_{j-1})^{\alpha}$, where α denotes an inflation factor. Estimate $I[g^{(m)}]$ via a local polynomial of order $l^* = l + 2$ and with h_d, j. Now, we obtain h_{j-1} by

$$h_j = \left(\frac{[m!]^2}{2m} \frac{(1-2\hat{d})}{\beta^2} \frac{(d_b - c_b)\hat{V}(1)}{I[\hat{g}^{(m)}]} \right)^{1/(2m+1-2\hat{d})} \cdot n^{(2\hat{d}-1)/(2m+1-2\hat{d})}. \qquad (10)$$

- **iv)** Repeat steps ii) and iii) until convergence or a given number of iterations has been reached and set $\hat{h}_{opt} = h_j$.

We propose to set the initial bandwidths to $h_0 = 0.1$ for $l = 1$ and $h_0 = 0.2$ for $l = 3$. Moreover, for $p = 3$ it is recommended to employ $c_b = 1 - d_b$ such that only 90% of all observations are used for estimating an interior point in order to reduce the boundary effect. For $l = 1$ all observations are used and hence $c_b = 1 - d_b = 0$. The bandwidth $h_{d.j}$ used for estimating $g^{(m)}$ is enlarged by means of an exponential inflation factor denoted by α. We have $\alpha = \alpha_{\text{opt}} = (2m+1-2d)/(2m+3-2d)$, $\alpha = \alpha_{\text{nai}} = (2m+1-2d)/(2m+5-2d)$ and $\alpha = \alpha_{var} = \frac{1}{2}$. Using α_{opt} results in bandwidth $h_{d,j}$ that minimizes the MSE of $\hat{I}[g^{(m)}]$ and consequently the rate of convergence of $\hat{h}_j$ is optimal. Whereas for α_{nai} the optimal rate of convergence is achieved for $\hat{m}^{(m)}$ and α_{var} ensures a stable selection of the

bandwidth. Moreover, we have $\alpha_{\text{var}} > \alpha_{\text{nai}} > \alpha_{\text{opt}}$ and $\alpha_{\text{nai}} \to \alpha_{\text{var}}$ as $d \to 0.5$. The choice of α depends on the underlying data, which is to be analysed. For a more detailed insight on inflation methods we refer the reader to Beran and Feng (2002a).

2.4 Data-driven estimation of g' and g''

The IPI for SEMIFARIMA models can also be applied to bandwidth selection for estimating $g^{(\nu)}$ with $\nu > 0$. In this paper, only the cases for $\nu = 1$ and $\nu = 2$ are discussed. The proposed IPI is now employed as a data-driven pilot method to obtain estimates for d, c_f and $h_{\nu,0}$ with order l_d, say. Estimation of $g^{(\nu)}$ is then carried out with $l = \nu + 1$ and $m = \nu + 2$. As previously, $g^{(m)}$), which is required for calculating $I[g^{(m)}]$ is estimated with order $l^* = l + 2$. The following two-stage procedure is proposed.

- **i)** In the first stage $\hat{d}$, $\hat{c}_f$, and $\hat{h}_{\text{opt}}$ are obtained by means of the main IPI-algorithm for estimating g with order $l_d = 1$ or $l_d = 3$.
- **ii)** Set $h_{\nu,0} = \hat{h}_{\text{opt}}$. Carry out an IPI-procedure as proposed above with fixed $\hat{c}_f$ and $\hat{d}$ in order to select a bandwidth for estimating $g^{(\nu)}$. Please note that (8) should be used.

Explicit formulas of the equivalent kernels for estimating $g^{(\nu)}$ at an interior point τ_t can be found in Müller (1988). The corresponding inflation factors are defined as previously and are determined by m and d.

3 Implementation in R

Based on the algorithms introduced in the previous section a R-package is developed, which is an extension of the already published *smoots* package. Hence, this package will be coined *smootslm*. The main functions are called *tsmoothlm* and *dsmoothlm* for estimating the trend and its derivatives, respectively, under presence of long-memory errors. Local polynomial estimation of $g^{(\nu)}$ and kernel smoothing of g are carried out by means of the functions *gsmooth* and *knsmooth*, which are implemented in the *smoots* package (see Feng et al., 2020).

In the following the function *tsmoothlm* is explained in more detail. The first argument y denotes the input time series. The second and third argument (*pmin* and *pmax*) are the minimum and maximum AR-order of the stochastic part Z_t in (3), respectively. Accordingly, the fourth and fifth argument (*qmin* and *qmax*) stand for the the minimum and maximum MA-order of Z_t. All four arguments can take the value 0, 1, 2, 3, 4 or 5 while $p_{\min} \leq p_{\max}$ and $q_{\min} \leq q_{\max}$. The optimal order is determined via BIC. The default setting is $p_{\min} = q_{\min} = p_{\max} = q_{\max} = 0$. The order of the polynomial for trend estimation is set via the argument p and the user can choose between 1 and 3, where $p = 1$ is the predefined option. The argument *mu* controls for the smoothness of the weight function. We have $\mu \in \{0, 1, 2, 3\}$, with $\mu = 1$ for the Epanechnikov kernel as default. Furthermore, the inflation factor α can be selected by the argument *InfR* with three different options, i.e. *"Opt"*, *"Nai"* and *"Var"*, which corresponds to $\alpha_{\text{opt}} = (2m + 1 - 2d)/(2m + 3 - 2d)$, $\alpha_{\text{nai}} = (2m+1-2d)/(2m+5-2d)$ and $\alpha_{var} = \frac{1}{2}$, respectively. The default setting for *InfR* is *"Opt"*. Moreover, the starting bandwidth h_0 can be set beforehand by the argument *bStart* with default $h_0 = 0.1$ for $p = 1$ and $h_0 = 0.2$ for $p = 3$. However, the choice of *bStart* should not affect the finally selected bandwidth if the IPI converges. Argument *bb* controls for boundary bandwidth. The default is $bb = 1$ meaning that the k-nearest neighbour method is applied, which results in a total bandwidth of $2\hat{h}$ at each observation point τ_t. For $bb = 0$ however, the total bandwidth is shortened at boundary points. By the argument *cb*, which is set to $cb = 0.05$ per default, the percentage of observations omitted for calculating $I[g^{(m)}] = \int_{c_b}^{1-d_b} [g^m(\tau)]^2 d\tau$ in (8) can be controlled. Additionally, the smoothing method with $\hat{h}_{\text{opt}}$ can be selected via the argument *method*. The user may choose between local polynomial regression ("*lpr*") and kernel regression ("*kr*"). However, originally kernel regression has been only incorporated in the *smoots* package as a benchmark to local polynomial regression.

For estimating $g^{(\nu)}$ the function *dsmoothlm* is applied. Please recall that here *tsmoothlm* is employed as a pilot method with minimum and maximum AR- and MA-order *pmin.p*, *pmax.p*, *qmin.p* as well as *qmax.p*, a local polynomial estimator of order *pp*, the inflation rate *InfR.p*, the kernel *mu.p* and a starting bandwidth *bStart.p* in order to obtain reasonable estimates for $\hat{d}$, $\hat{c_f}$ and $\hat{h}_{\nu,0}$ (see section 3.2). The options and default settings for these arguments are the same as for *tsmoothlm*. In addition to that, the order of the derivative to be estimated is set via the argument *nu* and can take the value 1 or 2 for the first and second derivative, respectively. Moreover, the argument *mu* controls the kernel

used for bandwidth selection after the pilot stage. Please note that the S3 methods (*print* and *plot*) implemented in the *smoots* package can be employed to the estimation results of the functions above. The output objects of *tsmoothlm* and *dsmoothlm* are basically lists containing input parameters and estimation results. Further detailed information on the functions are to be published in the users guideline of the *smootslm* package.

4 Application to different kinds of time series

In this and the following sections the SEMIFARIMA and ESEMIFARIMA are applied to four real data examples: *tempNH* (mean monthly temperature changes), *gdpGER* (GER GDP), *dax* (German stock index) and *vix* (CBOES volatility index). The *tempNH* data set has already been subject to an application example in Feng (2007), where the author employed the original version of the IPI for SEMIFAR models. The remaining three data sets have been already used by Feng et al. (forthcoming) and are implemented in the *smoots* package, which was recently published on the *CRAN* network.

4.1 Application to environmental data

The SEMIFARIMA model defined by (3) and (4) is applied to the time series of mean monthly Northern Hemisphere temperature changes (NHTM) from 1880 to 2018. The data is available at the website of the National Aeronautics and Space Administration (NASA). Bandwidth selection is carried out by means of the IPI for SEMIFARIMA models introduced in section 3. For model- and bandwidth selection the *tsmoothlm* function is applied, with $p = 1$, $pmin = qmin = 0$, $pmax = qmax = 3$ and $InfR =$ *"Opt"*. The remaining arguments are set on their default.

In Figure 1a) the fitted trend together with the observations is illustrated. The optimal bandwidth is $\hat{h}_{\text{opt}} = 0.165$ and a FARIMA $(0, \hat{d}, 0)$ has been selected following the BIC with $\hat{d} = 0.405$ implying strong long-range dependence in the temperature data. Apparently, the SEMIFARIMA captures the trend quite well. A clear upward trend can be observed approximately after 1970, which could be interpreted as an indicator for global warming. The trend-adjusted residuals are shown in Figure 1b) and first and second

derivative are depicted in Figures 1c) and 1d), respectively. Please note that for the estimation of derivatives the dependence structure has been estimated by pilot smoothing with order $pp = 3$. The derivatives match the features of the trend shown in Figure 1a) and provide further information about global temperature changes. For instance the slope of the first derivative indicates how strong the trend is increasing or decreasing. The intersections of the second derivative with the x-axis indicate a shift in the slope of $\hat{g}$.

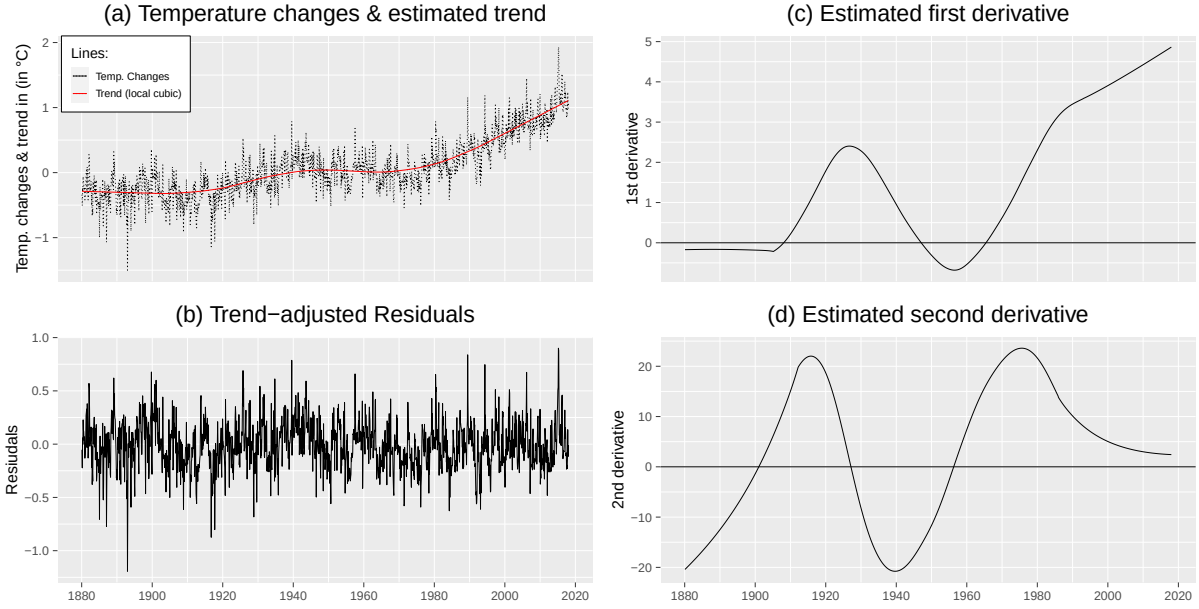

Figure 1: Estimated trend, residuals and the trend‘s derivatives for the NHTM series

4.2 Application to GDP data

A model which is commonly used in the field of macroeconomic research is the well-known log-linear growth model. Feng et al. (2020) have achieved a semiparametric local-linear extensions of this model by applying a Semi-ARMA to log-transformed GDP series. We follow this approach and additionally incorporate long memory by assuming that the log-transformed annually GER-GDP series from 1850 to 2016 follows a SEMI-FARIMA defined by (3) and (4). The data was obtained from the Maddison Project Database (2020). For this purpose the *tsmoothlm* function is employed, with $p = 1$, $pmin = qmin = 0$, $pmax = qmax = 3$ and $InfR =$ *"Opt"*. The remaining arguments are set on their default. We obtained an optimal bandwidth of 0.161 and a FARIMA $(2, \hat{d}, 2)$

with $\hat{d} = 0.342$, and $\hat{\phi}_1 = -0.319$, $\hat{\phi}_2 = 0.467$, $\hat{\psi}_1 = 1.421$ and $\hat{\psi}_2 = 0.336$ selected by the BIC for the residuals.

As a benchmark a kernel regression is carried out using the same bandwidth. Estimated trends together with log-gdp series are shown in Figure 2(a). At the interior both estimators are approximately equal. However, we can see that the kernel estimator clearly shows poor estimation quality at the boundaries which indicates that the local-linear estimator is to be preferred. The trend-adjusted residuals obtained by the local-linear approach are depicted in Figure 2(b). Moreover, the corresponding derivatives are illustrated in Figures 2(c) and 2(d), which reveal further information on the course of the German economy.

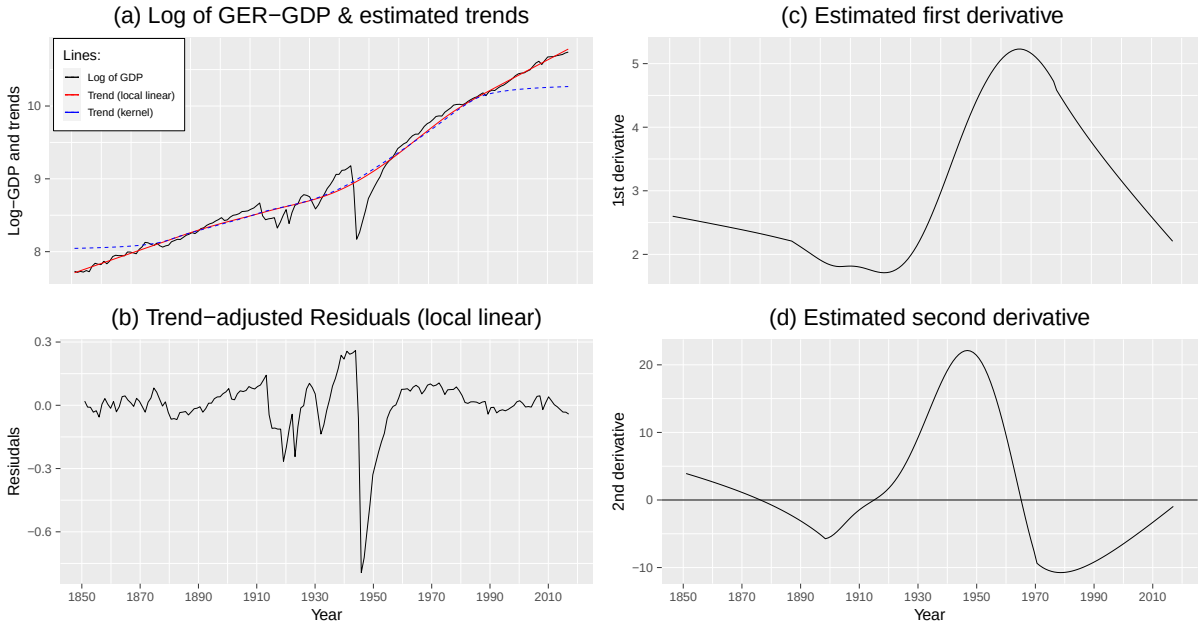

Figure 2: Estimated trend, residuals and the trend‘s derivatives for the GER-GDP series

5 Application to high-frequency data

The autoregressive conditional heteroscedasticity (ARCH) model proposed by Engle (1982) and its generalisation, the generalized ARCH (GARCH)) model, introduced by Bollerslev (1986) , is a well-known volatility process approach for modelling non-constant conditional variances. Feng (2004) found that return series often simultaneously exhibit conditional heteroskedasticity and a slowly changing scale. Under regular conditions a process

with conditional heteroskedasticity is covariance stationary, but a process with change in volatility is at best locally stationary. However, the majority of GARCH extensions are defined assuming a stationary return series. Based on his findings the author proposed the Semi-GARCH model by adding a smooth scale function to the standard GARCH model. Recently, Feng et al. (forthcoming) introduced the Semi-Log-GARCH model which is an extension of the Log-GARCH introduced by Pantula (1986), Geweke (1986) and Milhøj (1987). Moreover, Feng et al. (2020) and Letmathe et al. (forthcoming) proposed the FI-Log- and Semi-FI-Log-GARCH models, respectively. The FI-Log-GARCH is a special case of the FIAPARCH.

Let r_t^*, $t = 1, ..., n$ denote a return series with $E(r_t^*) = \mu_{r^*}$. We have

$$r_t = \sigma(\tau_t)\sqrt{h_t}\eta_t, \tag{11}$$

where $r_t = r_t^* - \mu_{r^*}$ are the centralized returns, $\sigma^2(x_t) > 0$ stands for a smooth scale function and η_t is defined as in (1). Let $\xi_t^2 = r_t^2/\sigma^2(\tau_t) = h_t\eta_t^2$ and $\alpha_{d,i} = \psi(B) - \phi(B)(1-B)^d$, where $\phi(B) = 1 - \sum_{i=1}^{p^*} \alpha_i B^i - \sum_{j=1}^{q} \beta_j B^j$ with $p^* = \max(p, q)$ and $\psi(B) = 1 + \sum_{j=1}^{q} \psi_j B^j = 1 - \sum_{j=1}^{q} \beta_j B^j$. Furthermore, α_i and β_j are the ARCH and GARCH coefficients, respectively. $\{\xi_t\}$ is assumed to follow a FI-Log-GARCH process given by

$$\ln h_t = \alpha_0 + \sum_{i=1}^{\infty} \alpha_{d,i} \ln \xi_{t-1}^2 + \sum_{j=1}^{q} \beta_j \ln h_{t-j}, \tag{12}$$

where α_0 is some constant. (11) and (12) together define a SEMI-FI-Log-GARCH. To ensure that our model is well defined we assume that $\xi_t \neq 0$ a.s. and $\text{var}\,(\xi_t) = 1$. Let $y_t = \ln r_t^2$, $g(\tau_t) = \ln \sigma^2(\tau_t) + \mu_{l\xi^2}$ and $Z_t = \ln \xi_t^2 - \mu_{l\xi^2}$, where $\mu_{l\xi^2} = E(\ln \xi_t^2)$. We can see that the log-transformation of the Semi-FI-Log-GARCH defined by (11) and (12) admits an additive model form $y_t = g(\tau_t) + Z_t$, which is a special case of Model (3). Moreover, let $\epsilon_t = \ln \eta^2 - \mu_{l\epsilon^2}$, with $\mu_{l\epsilon^2} = E(\ln \eta_t^2)$ and then we have $Z_t = \ln h_t + \epsilon_t - (\mu_{l\xi^2} - \mu_{l\epsilon^2})$. It was shown by Feng et al. (2020) that Z_t can be represented as a FARIMA(p^*, d, q) model given by

$$(1 - B)^d \phi(B)(Z_t) = \psi(B)\epsilon_t, \tag{13}$$

which is in turn a special case of Model (4). We see that the Semi-FI-Log-GARCH is equivalent to a SEMIFARIMA model with the restriction $p \geq q$. Subsequently, well

developed SEMIFARIMA algorithms are applicable for estimating $g(\tau_t)$ and Z_t. $\hat{\sigma}(\tau_t)$ can be obtained by $\hat{\sigma}(\tau_t) = \hat{C}_\sigma \exp[\hat{g}(\tau_t)/2]$, where $C_\sigma^2 = \text{var}\,(r_t/\exp[g(\tau_t)/2])$. C_σ can be estimated consistently by the scale-adjusted returns under the assumption that $E(\xi_1^4) < \infty$.

The Semi-FI-Log-GARCH is applied to the S&P500 series from 1990 to December 2020, which was downloaded from Yahoo Finance. Please note that it is possible to observe zero returns. Consequently, the log-transformation is not employable. As a remedy the squared returns are centralized, which is necessary anyway if the underlying return series has a non-zero mean. The centralized returns are depicted in Figure 3a). In Figure 3b) the log-transformed returns together with the estimated trends are shown. A local-linear and local cubic estimator are employed for trend estimation and are indicated by the blue and red line, respectively. For both estimators bandwidth selection by means of the IPI-algorithm was carried out with the same settings, i.e. the default for all arguments except $p = 3$ for local cubic and $p.max = q.max = 3$ for both approaches. The selected bandwidths obtained with a local linear and local cubic estimator are 0.094 and 0.155, respectively. As we can see in Figure 3b) the unconditional variances are clearly not time invariant. Moreover, both smoothing approaches deliver pleasing results. However, the fitted trend obtained via local linear regression is slightly over-smoothed. Consequently, the local cubic trend is used for further analysis. The conditional volatilities ($\hat{\sqrt{h}}$) shown in Figure 3c) are approximately stationary whereas this is not the case for the total volatilities ($\hat{\sigma(\tau_t)}\hat{\sqrt{h}}$) illustrated in Figure 3d). Furthermore, the following FARIMA$(1, \hat{d}, 1)$ with $\hat{d} = 0.294$, $\hat{\phi}_1 = 0.208$ and $\hat{\psi}_1 = 0.487$, is obtained from the residuals of the log-data. We see that the dependence structure of the errors is clearly affected by long-range persistence. Short-memory models could falsely capture long memory as short range dependence which may lead to poor forecasting accuracy.

6 The Semi-FI-Log-ACD model

Another well-known method for analysing non-negative financial time series is the autoregressive conditional duration (ACD) model introduced by Engle and Russell (1998). An extension of the ACD is the (type 1) Log-ACD$_1$ proposed by Bauwens et al. (2008) which can considered to be a squared version of the Log-GARCH. A fractionally integrated

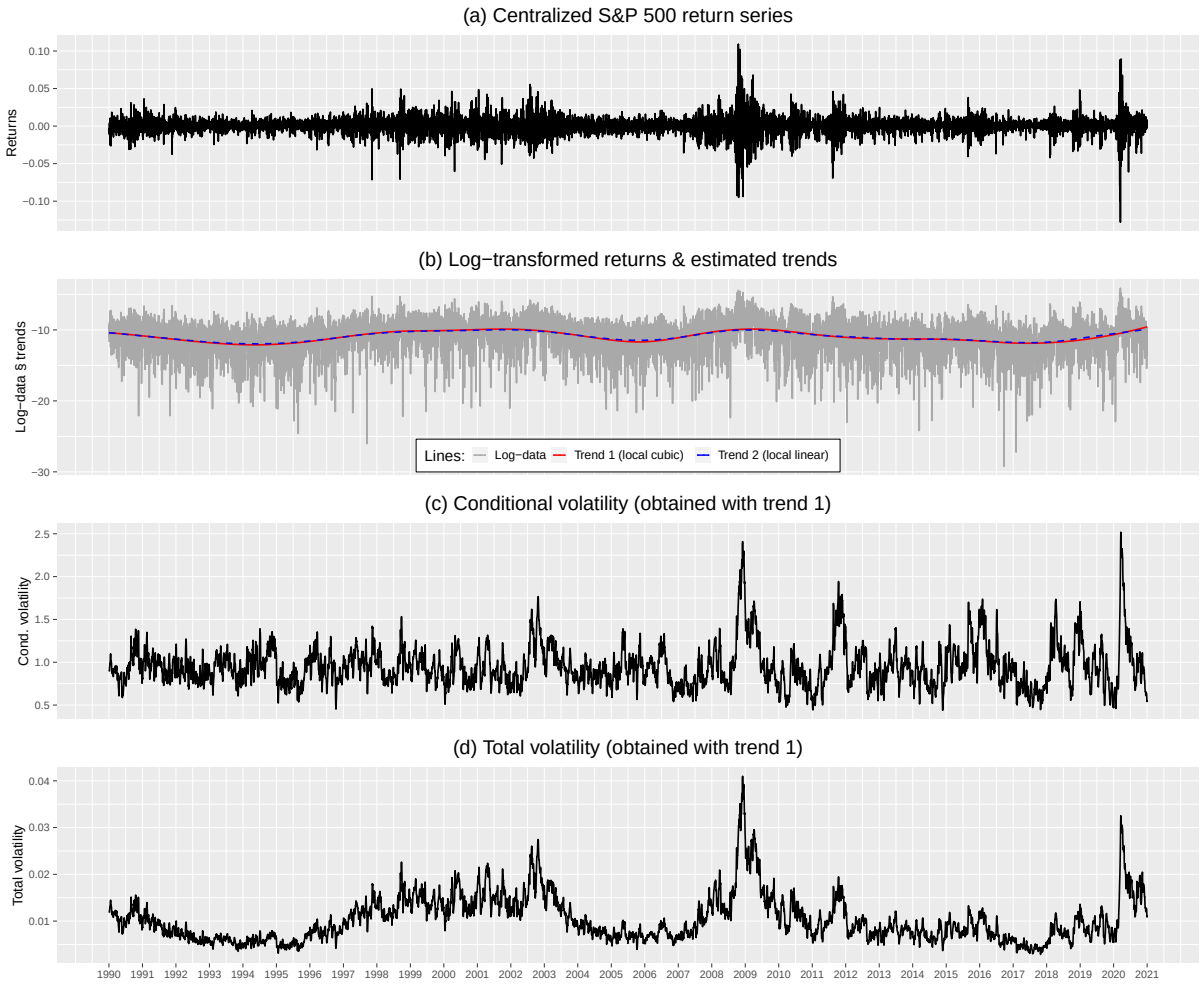

Figure 3: Estimation results of the Semi-FI-Log-GARCH for the SP500 series.

generalization of the Log-ACD was indicated by Beran et al. (2015) and subsequently Feng and Zhou (2015) proposed the FI-Log-ACD and its semiparametric extension the Semi-FI-Log-ACD. Let X_t, $t = 1, ..., n$ be some non-negative financial time series. A Semi-FI-Log-ACD for X_t is then defined by

$$X_t = s(\tau_t)\lambda_t e_t, \tag{14}$$

where $X^* = \lambda_t e_t$ follows a FI-Log-ACD, $s(\tau_t)$ is a smooth mean function, $\lambda_t > 0$ denotes the conditional mean and $e_t \geq 0$ are i.i.d. random variables. Furthermore, we assume that $E(\lambda_t) = E(e_t) = 1$. Moreover, the authors showed that the FI-Log-ACD (Semi-FI-Log-ACD) is equivalent to the EFARIMA (ESEMIFAR). Hence, the Semi-FI-Log-ACD corresponds to a squared Semi-FI-Log-GARCH defined by (11) and (12). By simply

replacing r_t^2, h_t and η_t^2 Semi-FI-Log-ACD can be estimated analogously. For a detailed derivation of the Semi-FI-Log-ACD we refer the reader to Feng and Zhou (2015).

In the following the Semi-FI-Log-ACD is applied to daily trading volume of the S&P500 from January 2000 to December 2020. The data was obtained from Yahoo finance as well. The original data X_t is displayed in Figure 4a) and it can be seen that the variation in the data is clearly time dependent. This indicates that using the log-trans Trend estimation is carried out with exactly the same settings as in the previous section except for $p = 3$ we have $InfR =$ "Nai". The selected bandwidths are 0.12 and 0.178. In Figure 3b) the log-transformed series, the local linear and local cubic trends are shown and are indicated by the black, red and blue (dashed) lines, respectively. Both estimators deliver very similar results but the local cubic approach seems to slightly over-fit the data. Therefore, the conditional and total means illustrated in Figures 4c) and 4d) as well as the residuals are obtained from the local linear estimates. From the residuals of the local linear estimator we obtain a FARIMA$(0, 0.446, 0)$ model with $\hat{d} = 0.446$ indicating strong long-range dependence in the data.

7 Concluding remarks

The paper at hand exemplifies the development of a supplementing R-package for the *smoots* package. The main feature of this package is the semi-parametric estimation under long-memory errors. In this regard an adapted version of the iterative plug in algorithm proposed by Beran and Feng (2002a) is introduced. Moreover, the implementation of this package is comprehensively described. The usage of two main functions is explained and illustrated by application to various non-stationary time series with long-memory. The estimation results are quite satisfactory. Further extensions of *smootslm* are the implementation of a forecasting procedure and the non-parametric estimation of the stochastic part of the model by means of e.g. a local Whittle-, GPH- or wavelet-estimator.

Acknowledgement: This paper was supported by the German DFG Project FE 1500/2-1. The sources of the used data are acknowledged in the context with corresponding references. We are grateful to Ms. Shujie Li, Mr. Bastian Schäfer and Mr. Dominik Schulz at Paderborn University, Germany, for helpful discussions and suggestions.

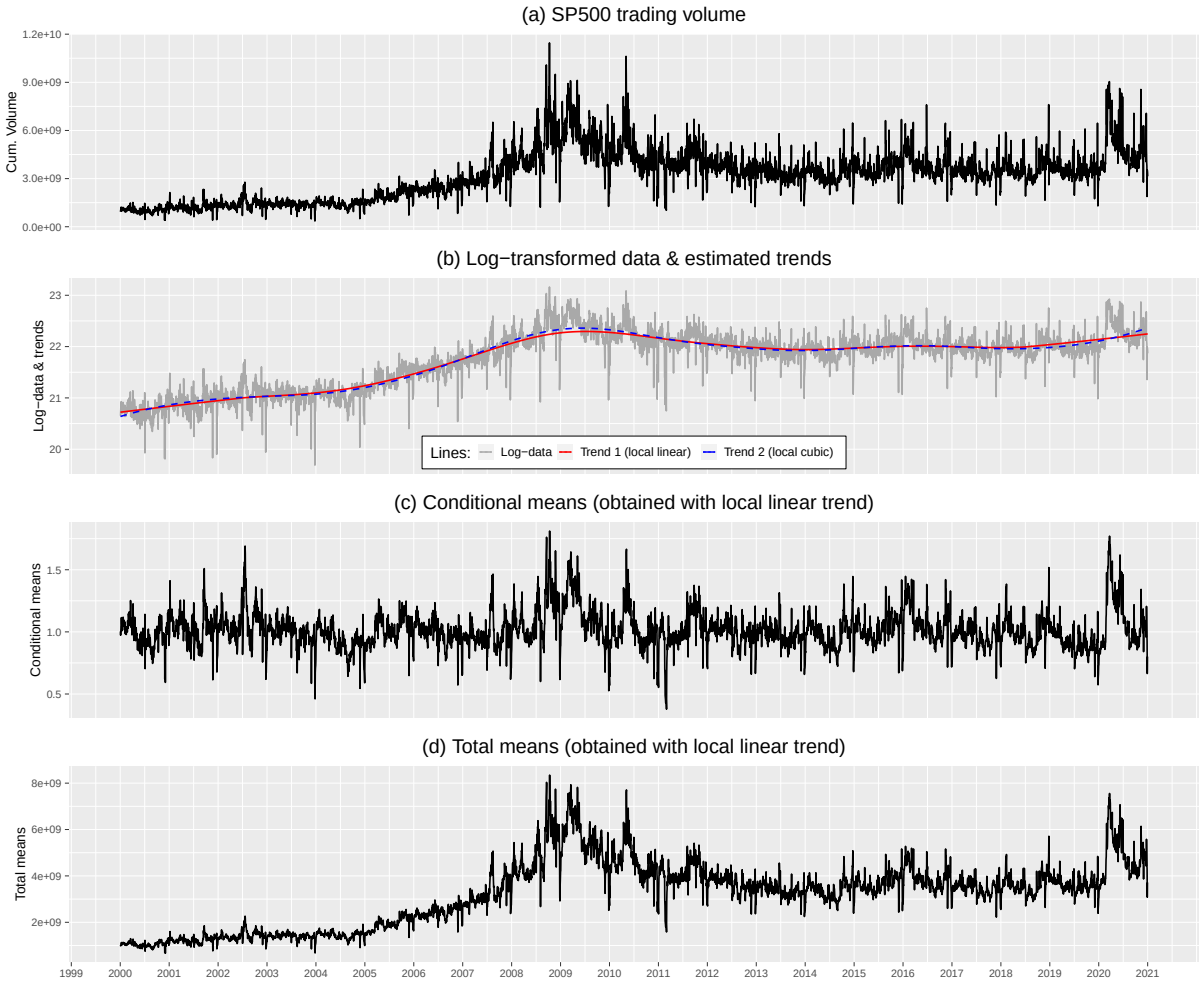

Figure 4: Estimation results of the Semi-FI-Log-ACD for the SP500 series.

References

Andersen, Torben G and Tim Bollerslev (1997). "Intraday periodicity and volatility persistence in financial markets". In: *Journal of empirical finance* 4.2-3, pp. 115–158.

Andersen, Torben G, Tim Bollerslev, and Steve Lange (1999). "Forecasting financial market volatility: Sample frequency vis-a-vis forecast horizon". In: *Journal of empirical finance* 6.5, pp. 457–477.

Bauwens, Luc, Fausto Galli, and Pierre Giot (2008). "Moments of the Log-ACD model". In: *Quantitative and Qualitative Analysis in Social Sciences* 2, pp. 1–28.

Bauwens, Luc and Pierre Giot (2000). "The logarithmic ACD model: an application to the bid-ask quote process of three NYSE stocks". In: *Annales d'Economie et de Statistique*, pp. 117–149.

Beran, Jan and Yuanhua Feng (2002a). "Iterative plug-in algorithms for SEMIFAR models—definition, convergence, and asymptotic properties". In: *Journal of Computational and Graphical Statistics* 11.3, pp. 690–713.

– (2002b). "Local polynomial fitting with long-memory, short-memory and antipersistent errors". In: *Annals of the Institute of Statistical Mathematics* 54.2, pp. 291–311.

– (2002c). "SEMIFAR models—a semiparametric approach to modelling trends, long-range dependence and nonstationarity". In: *Computational Statistics & Data Analysis* 40.2, pp. 393–419.

Beran, Jan, Yuanhua Feng, and Sucharita Ghosh (2015). "Modelling long-range dependence and trends in duration series: an approach based on EFARIMA and ESEMIFAR models". In: *Statistical Papers* 56.2, pp. 431–451.

Beran, Jan and Dirk Ocker (1999). "SEMIFAR forecasts, with applications to foreign exchange rates". In: *Journal of Statistical Planning and Inference* 80.1-2, pp. 137–153.

Beran, Jan et al. (2013). "Limit theorems". In: *Long-Memory Processes*. Springer, pp. 209–384.

– (2016). *Long-Memory Processes.* Springer.

Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3, pp. 307–327.

Cotter, John (2005). "Uncovering long memory in high frequency UK futures". In: *The European Journal of Finance* 11.4, pp. 325–337.

Ding, Zhuanxin and Clive WJ Granger (1996). "Modeling volatility persistence of speculative returns: a new approach". In: *Journal of econometrics* 73.1, pp. 185–215.

Ding, Zhuanxin, Clive WJ Granger, and Robert F Engle (1993). "A long memory property of stock market returns and a new model". In: *Journal of empirical finance* 1.1, pp. 83–106.

Engle, Robert (2002). "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models". In: *Journal of Business & Economic Statistics* 20.3, pp. 339–350.

Engle, Robert F (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica: Journal of the Econometric Society*, pp. 987–1007.

Engle, Robert F and Jeffrey R Russell (1998). "Autoregressive conditional duration: a new model for irregularly spaced transaction data". In: *Econometrica*, pp. 1127–1162.

Feng, Yuanhua (2004). "Simultaneously modeling conditional heteroskedasticity and scale change". In: *Econometric Theory* 20.3, pp. 563–596.

– (2007). "On the asymptotic variance in nonparametric regression with fractional time-series errors". In: *Nonparametric Statistics* 19.2, pp. 63–76.

Feng, Yuanhua and Chen Zhou (2015). "Forecasting financial market activity using a semiparametric fractionally integrated Log-ACD". In: *International Journal of Forecasting* 31.2, pp. 349–363.

Feng, Yuanhua et al. (2020). *Fractionally integrated Log-GARCH with application to value at risk and expected shortfall*. Tech. rep. Paderborn University, CIE Center for International Economics.

Gasser, Theo, Alois Kneip, and Walter Köhler (1991). "A flexible and fast method for automatic smoothing". In: *Journal of the american statistical association* 86.415, pp. 643–652.

Gasser, Theo and Hans-Georg Müller (1979). "Kernel estimation of regression functions". In: *Smoothing techniques for curve estimation*. Springer, pp. 23–68.

Geweke, John (1986). "Comment". In: *Econometric Reviews* 5.1, pp. 57–61.

Karanasos, M (2008). "The statistical properties of exponential ACD models". In: *Quantitative and Qualitative Analysis in Social Sciences* 2.1, pp. 29–49.

Letmathe, Sebastian, Yuanhua Feng, André Uhde, et al. (2021). *Semiparametric GARCH models with long memory applied to Value at Risk and Expected Shortfall*. Tech. rep. Paderborn University, CIE Center for International Economics.

Milhøj, Anders (1987). "A conditional variance model for daily deviations of an exchange rate". In: *Journal of Business & Economic Statistics* 5.1, pp. 99–103.

Müller, Hans-Georg (1988). "Longitudinal Data and Regression Models". In: *Nonparametric Regression Analysis of Longitudinal Data*. Springer, pp. 6–14.

Pantula, Sastry G (1986). "Modeling the persistence of conditional variances: a comment". In: *Econometric Reviews* 5, pp. 79–97.