Nonparametric regression with long memory using R

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

June 9, 2020

Abstract

A spatial SEMIFAR model is defined by introducing a smoothing component into the spatial FARIMA model. The application of this model to non-negative financial data via log-transformation is hence called spatial ESEMIFAR. This model is analyzed by partially applying the ESEMIFAR model to univariate financial time series, namely on all observations at a given trading day and on observations of all days at a given trading time. After elimination of a deterministic trend conditional fluctuations in both dimensions are further investigated. The estimation of the nonparametric component can be simply carried out through SEMIFAR-algorithms. The appropriateness of our approach is then investigated and illustrated by the application to financial high-frequency data. Trend estimation and model selection is implemented with the statistical software R.

Keywords: S-FARIMA, S-SEMIFAR, long memory, random fields, high-frequency data, local polynomial

JEL Codes: C14, C51

1 Introduction

Literature research and model research required.

In many areas of research data are observed spatially, depending on two separate dimensions in a lattice. In recent years one can observe more frequently some sort of apparent memory in the decay of spatial correlations to depend and change over its direction within the spatial process. For instance, long-memory in the sense of slowly decaying autocorrelations in (high frequency) financial data across trading time and trading day produces a random field on a lattice in both dimensions simultaneously. Beran, Feng, and Ghosh (2015) state that daily average trade duration data has often shown long memory with a clear non zero mode. Therefore a log-normal conditional distribution is suggested. The simplest approach to model long range dependence in a positive valued time series is to take the exponential of a linear long memory process such as FARIMA leading to stochastic volatility models. Due to the long range dependence there is an unobservable latent process which makes the estimation and interpretation of the fitted parameters very challenging.

The SEMIFAR and ESEMIFAR models introduced by Beran and Feng (2002c) and Beran, Feng, and Ghosh (2015) are designed for simultaneous modeling of stochastic trends, deterministic trends and stationary short- and long-memory components in a time series such that the trend generating mechanisms can be distinguished.

2 The models

2.1 The SEMIFAR

A process Y_t is said to follow a SEMIFAR model, introduced by Beran (1999) if there exists an integer $m \in \{0, 1\}$ and a fraction $\delta \in (-0.5, 0.5)$ such that

$$\phi(B)(1-B)^{\delta}\{(1-B)^{m}Y_t - g(x_t)\} = \epsilon_t, \tag{1}$$

where $\phi(x) = 1 - \sum_{j=1}^{p} \phi x^j$ is a polynomial with all roots outside the unit circle, ϵ_t are iid normal with $E(\epsilon_t) = 0$, $\text{var}(\epsilon_t) = \sigma_\epsilon^2$, $x_t = t/n$ with $t \in \mathbb{Z}$, B is the backshift operator and $g : [0, 1]$ is a nonparametric smooth trend function. The fractional differencing parameter δ was introduced by Granger and Joyeux (1980) and Hosking (1981) and is defined by

$$(1 - B)^\delta = \sum_{k=0}^{\infty} b_k(\delta) B^k, \tag{2}$$

with

$$b_k(\delta) = (-1)^k \binom{\delta}{k} = (-1)^k \frac{\Gamma(\delta + 1)}{\Gamma(k + 1)\Gamma(\delta - k + 1)}. \tag{3}$$

Considering the autocovariances $\gamma(k) = \text{cov}(Y_t, Y_{t+k})$, Y_t incorporates long memory if the spectral density given by

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{ik\lambda} \gamma(k) \tag{4}$$

exhibits a pole at the origin of the frequency spectrum such that

$$f(\lambda) \sim c_f |\lambda|^{-2\delta}, \quad (\text{as } \lambda \to 0), \tag{5}$$

where $c_f > 0$ and "$\sim$" stands for the ratio of both sides converging one. Then, for $k \to \infty$ the autocovariances $\gamma(k)$ are proportional to $k^{2\delta-1}$ and hence yield an infinite sum. We can distinguish between three temporal dependency structures. The process $Z_t = \{(1 - B)^m Y_t - g(x_t)\}$ has long memory for $\delta > 0$ with $\sum_{k=-\infty}^{\infty} \gamma_U(k) = \infty$, short memory for $\delta = 0$ with $\sum_{k=-\infty}^{\infty} \gamma_U(k) < \infty$ and is anitpersistent for $\delta < 0$ with $\sum_{k=-\infty}^{\infty} \gamma_U(k) = 0$ frequently reversing itself. Based on model (1) Beran and Feng (2002c) proposed an adapted version of a data-driven algorithm already introduced in Beran (1997) by replacing an estimate of the constant mean with a kernel estimate of g defined by

$$\hat{g}(x) = \frac{1}{nh} \sum_{t=1}^{n} K(\frac{x - x_t}{h}) Y_t, \tag{6}$$

where $h > 0$ denotes the bandwidth, $x \in [0, 1]$ and $K(\cdot)$ is a symmetric polynomial kernel (see e.g. Gasser and Müller 1979).

A comprehensive application of the SEMIFAR to finanical time series data was carried out by

2.2 The ESEMIFAR (Anwendung auf Log Daten, Eigenschaften Log normal ...)

Let Y_t be an equidistant and nonparametric additive regression model of the form

$$Y_t = g(x_t) + Z_t, \tag{7}$$

with standardized time $x_t = t/n$, a nonparametric smooth trend function $g : [0, 1]$ and some stationary process Z_t which is assumed to be fractionally differenced such that $(1 - B)^\delta Z_t = U_t$ holds, where $U_t = Z_t$ for short memory with $\delta = 0$. Therefore Z_t may be called a fractional ARIMA or FARIMA(p,δ,q) process capturing persistent dependency and is given by

$$(1 - B)^\delta \phi(B) Z_t = \psi(B)\epsilon_t, \tag{8}$$

where the fractional memory parameter $\delta \in (-0.5, 0.5)$, B as backshift operator, $\phi(B) = 1 - \sum_{k=1}^{p} \phi_k B^k$ and $\psi(B) = 1 + \sum_{k=1}^{q} \psi_k B^k$ are the characteristic autoregressive and moving average polynomials for short memory and $\epsilon_t \, (t \in \mathbb{Z})$ are iid. random variables with $E(\epsilon_t = 0)$ and $var(\epsilon_t = \sigma_\epsilon^2)$.

A time series Y_t in (7) including a smooth component and a stationary persistent dependence structure can be described by a SEMIFAR(p,δ,0) process (Beran, Bhansali, and Ocker (1998) and Beran and Feng (2002c)), since Y_t (with $q = 0$) can be modeled such that

$$\phi(B)(1 - B)^\delta \{(1 - B)^m Y_t - g(x_t)\} = \epsilon_t, \tag{9}$$

using an additive approach with nonparametric regression and long memory as a semiparametric extension of (8). Here $m \in \{0, 1\}$ gives the smallest possible integer for the estimation of differenced data. A semiparametric process as in (9) has long memory if (4) and (5) are satisfied which requires the autocovariances $\gamma(k)$ to behave proportional to $k^{2\delta-1}$ producing an infinite sum of autocovarinces for $k \to \infty$. Since the trend $g(x_t)$ is non parametric and not assumed to follow a particular form, (9) includes stochastic trends for $m > 0$ where the differenced process can show short- or long memory components, and deterministic trends for $m = 0$ with stationary short- or long memory components. Hence $g(x_t)$ is allowed to be a constant (e.g. the mean) or a smooth function with a stochastic and/or a deterministic trend. Consequently under $m = 0$ a continuous trend

plus stationary noise admits a spurious trend, whereas for $m > 0$ and $g \neq 0$ the trend is generated simultaneously by a deterministic and a stochastic trend. Then the differenced process without the trend is a stationary fractional ARIMA process (8). Further reading as to the spectral density representation for the memory of a process can be found in Mandelbrot (1983), Cox (1984), Kunsch (1987), Hampel (1987), and Beran (1995). In particular, for long memory in random fields view Angulo, Ruiz-Medina, and Anh (2000) and Anh, Angulo, and Ruiz-Medina (1999).

For the estimation of a smooth $g^{(\nu)}$ in (7) a local polynomial approach is investigated by Ruppert and Wand (1994), Wand and Jones (1994), Fan and Gijbels (1995), and Beran and Feng (2002b). A polynomial approximation of the unknown function $g(x)$ is used for x in neighborhood of a point x_t and is $(p+1)$ times differentiable. The local approximation of $g(x)$ is following Beran et al. (2016) and Beran and Feng (2002b) given by the p order polynomial

$$g(x_t) \approx g(x) + g^{(1)}(x)(x_t - x) + \ldots + g^{(p)}(x)\frac{(x_t - x)^p}{p!} + R_p, \tag{10}$$

with a remainder term R_p. Applying a polynomial regression locally for the observations $y_1, \ldots, y_n$ and using (**??**), the νth derivative of the estimator of $g^{(\nu)}(x)$ are given by $\hat{g}^{(\nu)}(x) = \nu!\hat{\beta}_\nu$ where the fixed coefficients $\beta_\nu = \beta_\nu(x) = g^{(\nu)}(x)/\nu!$ with $\nu = 0, 1, 2, \ldots, p$ such that (**??**) becomes $g(x_t) = \sum_{\nu=0}^{p}(x_t - x)^\nu \beta_\nu$. Assume the kernel K to be a symmetric density with compact support $[-1, 1]$ such that (see Gasser and Müller (1984))

$$K(x) = \sum_{l=0}^{r} \alpha_l x^{2l}, \quad (|x| \leq 1) \tag{11}$$

with $K(x) = 0$ for any point $|x| > 1$, $r \in \{0, 1, 2, \ldots\}$ and the integral of K between $(-1, 1)$ yields one. Then the estimator $\hat{g}^{(\nu)}(x)$ with $(\nu \leq p)$ is obtained when the locally weighted sum of squared residuals is minimized such that

$$Q(x) = \sum_{t=1}^{n} \left\{ y_t - \sum_{j=0}^{p} \beta_j (x_t - x)^j \right\}^2 K\left(\frac{x_t - x}{h}\right) \Rightarrow min, \tag{12}$$

with h as bandwidth and $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ and the weight term $K((x_t - x)/h)$ limits

the farthest neighbor for the polynomial fit for x. The matrix representation of (12) is

$$Q(x) = (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{K}(x)(\mathbf{y} - \mathbf{X}\beta), \tag{13}$$

with

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{bmatrix},$$

and

$$\mathbf{K} = \begin{bmatrix} K\left(\frac{x_1 - x}{h}\right) & 0 & \cdots & 0 \\ 0 & K\left(\frac{x_2 - x}{h}\right) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & K\left(\frac{x_n - x}{h}\right) \end{bmatrix},$$

leading to a solution of (12) and (13) and can according to Feng (2004) be written as

$$\begin{aligned} \hat{g}^{(\nu)}(x) = \nu! \hat{\beta}_\nu &= \nu! e_{\nu+1}^T (\mathbf{X}^T \mathbf{K} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{K} \mathbf{y} \\ &= \{\mathbf{w}^\nu(x)\}^T \mathbf{y}, \end{aligned} \tag{14}$$

the p-th order local polynomial estimator $\hat{g}^{(\nu)}$ with $\mathbf{y} = (y_1, \ldots, y_n)^T$, the weighting system $\{\mathbf{w}^\nu(x)\}^T = (w_1^\nu, \ldots, w_n^\nu)^T = \nu! e_{\nu+1}^T (\mathbf{X}^T \mathbf{K} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{K}$ and using the unit vector $e_\nu = (e_{1,\nu}, \ldots, e_{(p+1),\nu})^T$ with $\nu = 1, \ldots, p+1$ where $e_{\nu,\nu} = 1$, $e_{i,\nu} = 0$ for $i \neq \nu$. The weighting system for any t is nonzero only if $|x_t - x| \leq h$ and therefore all interior points within $[h, 1-h]$ share the same w^ν and are independent of the errors dependence structure. Furthermore the weights have under any design the properties

$$\sum_{t=1}^{n} w^\nu (x_t - x)^\nu = \nu! \quad \text{and} \quad \sum_{t=1}^{n} w^\nu (x_t - x)^l = 0, \tag{15}$$

for $l = 0, \ldots, p$ where $j \neq \nu$. Since (15) yields 0 except for $l = \nu$, $\hat{g}^{(\nu)}$ is unbiased if the polynomial is of maximal order p. This relation is also true for not equidistant In consequence a local polynomial estimator is boundary corrected for odd $p-\nu$ as the order of the bias for interior and boundary points is the same whereas kernel estimators only approximate the properties and encounter boundary point problems. The calculation of the bias of $\hat{g}^{(\nu)}(x)$ for $p-\nu$ odd and even is omitted and can be seen in Beran et al. (2016).

3 Data-driven estimation procedure

A spatial extension of the SEMIFAR model in (7) to a higher dimensional space is achieved by running Y_t over a second index j such that a random field on a lattice originates along the trading time j and the trading day t. Let $Y_{t,j}$ be an equidistant, spatial and nonparametric regression model of the form

$$Y_{t,j} = g(x_t, \tau_j) + Z_{t,j}, \tag{16}$$

with a non negative random field, rescaled $x_t = t/n_1$, $\tau_j = j/n_2$ and a trend $g(\cdot,\cdot) : [0,1] \to \mathbb{R}$ which is a smooth lattice process. The multivariate $Z_{t,j}$ in (16) is a stationary lattice process and assumed to follow a zero mean S-FARIMA model which was introduced by Beran, Ghosh, and Schell (2009) in order to account for the two-dimensional anisotropic memory of a process and is expressed by

$$(1-B_1)^{\delta_1}(1-B_2)^{\delta_2}\phi_1(B_1)\phi_2(B_2)Z_{t,j} = \psi_1(B_1)\psi_2(B_2)\epsilon_{t,j}, \tag{17}$$

where the (long) memory parameters $\delta_1, \delta_2 \in (-0.5, 0.5)$ cover both directions in the lattice with horizontal δ_1 and vertical δ_2. The $\epsilon_{t,j}\,(t,j \in \mathbb{Z})$ are iid random variables with $E(\epsilon_{t,j}) = 0$, $var(\epsilon_{t,j}) = \sigma_\epsilon^2$ and the backshift operators produce $B_1 X_{t,j} = X_{t-1,j}$ and $B_2 X_{t,j} = X_{t,j-1}$. In direction of the column Y_{t,j_0} with $t = 1,\ldots,n_1$ a specific trading time j_0 across all days t is considered. Then, for the row $Y_{t_0,j}$ with $j = 1,\ldots,n_2$ all observations j within one trading day t_0 are considered. The stationary series $Z_{t,j}$ features short memory with finite summed autocovariances $\sum \gamma_Z(k) = c$ for $0 < c < \infty$. The autoregressive and moving average components for each direction are given by

$$\begin{aligned} \phi_1(B_1) &= 1 - \sum_{k=1}^{p_1} \phi_{1,k} B_1^k, \quad \phi_2(B_2) = 1 - \sum_{k=1}^{p_2} \phi_{2,k} B_2^k, \\ \psi_1(B_1) &= 1 + \sum_{k=1}^{q_1} \psi_{1,k} B_1^k, \quad \psi_2(B_2) = 1 + \sum_{k=1}^{q_2} \psi_{2,k} B_2^k, \end{aligned} \tag{18}$$

and the roots of the polynomials ϕ_1, ϕ_2 and ψ_1, ψ_2 are outside the unit circle with $\delta_1, \delta_2 \in (0, 0.5)$ such that $Z_{t,j}\,(t, j \in \mathbb{Z})$ is stationary and invertible allowing the notation

$$\begin{aligned} \epsilon_{t,j} &= (1-B_1)^{\delta_1}(1-B_2)^{\delta_2}\left[\frac{\phi_1(B_1)}{\psi_1(B_1)}\frac{\phi_2(B_2)}{\psi_2(B_2)}\right] Z_{t,j} \\ &= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} b_r(\eta_{\text{col}})b_s(\eta_{\text{row}})Z_{t-r,j-s}, \end{aligned} \tag{19}$$

indicating the linear row and column filter. One main characteristic of (17) is, that there is a fractional ARIMA type dependence structure involved. In consequence, the stationary random field $Z_{t,j}$ has a spectral density of the form

$$\begin{aligned} f(\lambda_1, \lambda_2) &= \frac{\sigma_\epsilon^2}{4\pi^2}\left|1-e^{-i\lambda_1}\right|^{-2\delta_1}\left|1-e^{-i\lambda_2}\right|^{-2\delta_2}\left|\frac{\psi_1(e^{-i\lambda_1})}{\phi_1(e^{-i\lambda_1})}\right|^2\left|\frac{\psi_2(e^{-i\lambda_2})}{\phi_2(e^{-i\lambda_2})}\right|^2 \\ &= \sigma_\epsilon^2 f_1(\lambda_1) f_2(\lambda_2), \end{aligned} \tag{20}$$

where λ_1, λ_2 are frequencies and $f_i (i = 1, 2)$ are the spectral densities of a spatial fractional ARIMA model with innovation variance one, p_i, δ_i, q_i and the moving average and autoregressive polynomials ϕ_i and ψ_i. Moreover notice the area A_{pole} specifying the pole in the origin of the frequency spectrum with $\{\lambda_1, \lambda_2\}$ and its borders by

$$\begin{aligned} A_{pole} &= \left\{\lambda \in [-\pi, \pi]^2 : \lambda_1 = 0, \lambda_2 \in [-\pi, \pi]\right\} \cup \left\{\lambda \in \mathbb{R}^2 : \lambda_1 = [-\pi, \pi], \lambda_2 = 0\right\} \\ &= \{\lambda : \lambda_1 + \beta\lambda_2 = 0 \,\text{with}\, \beta = 0\} \cup \{\lambda : \beta\lambda_1 + \lambda_2 = 0 \,\text{with}\, \beta = 0\} \end{aligned} \tag{21}$$

which holds only if $\delta_1, \delta_2 > 0$ and allows easy application to irregularly shaped sampling areas different from a square grid which is a quite mild assumption. Once $Z_{t,j}$ can be described by a S-FARIMA model as in (17), $E(Z_{t,j}) = 0$ and $var(Z_{t,j}) = \sigma_Z^2$, then $Y_{t,j}$ in (16) incorporates a stationary persistent dependence structure in addition to the smooth component g. Following Beran, Ghosh, and Schell (2009) $Y_{t,j}$ can be called a spatial SEMIFAR (S-SEMIFAR) process. It is modeled using nonparametric regression with long memory (Hall and Hart, 1990). The parameter vector to be determined by for instance the maximum likelihood method is $\theta = (\sigma_\epsilon^2, \eta^T)$, where $\eta = (\eta_{\text{row}}^T, \eta_{\text{col}}^T)$ and

$$\eta_{\text{row}} = (\delta_1, \phi_{1,1}^T, \ldots \phi_{1,p_1}^T, \psi_{1,1}^T, \ldots \psi_{1,q_1}^T)^T, \quad \eta_{\text{col}} = (\delta_2, \phi_{2,1}^T, \ldots \phi_{2,p_2}^T, \psi_{2,1}^T, \ldots \psi_{2,q_2}^T)^T, \tag{22}$$

indicating the estimation of the spatial $Y_{t,j}$ with two separate long memory parameters in j and t direction. SEMIFAR algorithms for one dimensional processes can be applied separately. The literature offers some alternative definitions for spatial models with long memory introduced by Lavancier (2007), Lavancier (2008), and Guo, Lim, and Meerschaert (2009). If the log transformation is applied, the FARIMA model (17) can be used to exhibit long memory also in a multiplicative model for non-negative financial data. Analogously, the SEMIFAR model was applied to the log transformation of such data in order to model long memory and a multiplicative trend at the same time (Chen, Yu, and Zivot, 2012). In order to investigate the conditional volatility of a series on the financial market its nonparametric components must first be estimated and then eliminated. In this way further possible random effects can be determined. The estimation of the trend on a trading day t_0 including all selected 511 trading minutes j is given by

$$\hat{g}(x_{t_0}, \tau) = \frac{\sum_{j=1}^{n_2} K_2\left(\frac{\tau_j - \tau}{h_2}\right) Y_{t_0,j}}{\sum_{j=1}^{n_2} K_2\left(\frac{\tau_j - \tau}{h_2}\right)}, \quad t_0 = 1, \ldots, n_1, \tag{23}$$

and estimation of the trend at a fixed trading time j_0 from all trading days t is given by

$$\hat{g}(x, \tau_{j_0}) = \frac{\sum_{t=1}^{n_1} K_1\left(\frac{x_t - x}{h_1}\right) Y_{t,j_0}}{\sum_{t=1}^{n_1} K_1\left(\frac{x_t - x}{h_1}\right)}, \quad j_0 = 1, \ldots, n_2, \tag{24}$$

where $K_1(u_1)$, $K_2(u_2)$ are kernel functions and b_1, b_2 are the associated bandwidths used for the application of the local polynomial regression. While the estimated trends in (23) show the daily patterns of the financial market on all trading days, the estimation in (24) gives the long range dynamics at different trading times. The function $g(\cdot, \cdot)$ in (7) can be estimated by bivariate nonparametric regression to read in Ruppert and Wand (1994), Härdle and Müller (1997), and Scott (2015).

4 IPI-algorithm for the regression and its derivatives

Text...

To avoid a full search for each trial value of $d \in G$ all parameters are recommended to be estimated directly from the residuals, using the maximum likelihood method (Beran and

Feng, 2002c), whereas m and p will be determined by the original data. This technique has shown to be much faster in computation compared to a full search approach. Using the integral of the trend(-derivative) without borders and the following notation (view Beran and Feng (2002b))

$$V_n(\theta, h) = (nh)^{-1-2\delta} \sum_{t,j=nx-nh}^{nx+nh} K\left(\frac{x - x_j}{h}\right) K\left(\frac{x - x_t}{h}\right) \gamma(t-j), \tag{25}$$

$$I(g^{(k)}) = \int_{\Delta}^{1-\Delta} \left(g^{(k)}(x)\right)^2 d\tau \ \neq 0, \tag{26}$$

and the kernel constant as

$$I(K) = \int_{-1}^{1} u^2 K(u) du. \tag{27}$$

Then, for $\delta \in (-0.5, 0.5)$, $n \to \infty$ and $h \to 0$, such that $nh \to \infty$ we have a bias of

$$E\left[\hat{g}^{(\nu)}(x) - g^{(\nu)}(x)\right] = h_n^{(k-\nu)} \frac{g^{(k)}(x) I(K)}{k!} + o(h_n^{(k-\nu)}) \tag{28}$$

uniformly in $\Delta < x < 1 - \Delta$. For increasing sample size

$$\lim_{n\to\infty} V_n(\theta, h_n) = V(\theta) \tag{29}$$

with constant and finite $0 < V(\theta) < \infty$. The uniform variance for the interior

$$(nh_n)^{1-2\delta} h_n^{2\nu} \mathrm{var}(\hat{g}^{(\nu)}(x)) = V(\theta) + o(1). \tag{30}$$

Hence, the MISE of $\hat{g}^{(\nu)}$ in $[\Delta, 1-\Delta]$ to be minimized is given by

$$\begin{aligned} &\int_{\Delta}^{1-\Delta} E\left\{\left[\hat{g}^{(\nu)}(x) - g^{(\nu)}(x)\right]^2\right\} dx \\ &\quad = \mathrm{MISE}_{asympt}(n, h_n) + o(\max(h^{2(k-\nu)}, (nh_n)^{2\delta-1} h^{-2\nu})) \\ &\quad = h_n^{2(k-\nu)} \frac{I(g^{(k)}) I^2(K)}{k!} + (nh_n)^{2\delta-1} h^{-2\nu} V(\theta) \\ &\qquad + o(\max(h^{2(k-\nu)}, (nh)^{2\delta-1} h^{-2\nu})), \end{aligned} \tag{31}$$

where ν gives the level of the derivative, p is the order of the polynomial and $k = p + 1$ gives the minimal differentiability of g on $[0, 1]$. Lower polynomial order with odd $p - \nu$

includes the $(\nu+1)$th standard local polynomial estimator $\hat{g}^{(\nu)}$. Explicit formulas for V can be given for short memory with $\delta = 0$, long memory with $\delta > 0$ and antipersistence with $\delta < 0$ and can be found in Hall and Hart (1990) and Beran and Feng (2002a). δ and V may be estimated consistently by standard maximum likelihood methods, where $V(0) = \int_{-1}^{1} K^2(u)du = R(K)$. In case $\delta \neq 0$ a new kernel density function has to be fitted to the chosen kernel and its order. The resulting $V(\delta)$ depends on the kernel function and is part of the asymptotic variance for $\delta \in (-0.5, 0.5)$. Let $K(u)$ be a polynomial kernel as in (11) on $[-1, 1]$ and $m = 2\,(m = 3)$ for k even (odd). Then, following Feng (2003) we have

$$V(\delta) = Vc \left[\sum_{(l-k)even} \alpha_l^2 T_{l,l} + 2 \sum_{\substack{l \geq m' \\ (l-k)even}} \sum_{\substack{m<l \\ (m-k)even}} \alpha_l \alpha_m T_{l,m} \right], \tag{32}$$

with $l, m = 0, 1, \ldots, r$ in (11) such that $(l-k)$ and $(m-k)$ are both even with

$$\begin{aligned} T_{l,m} &= \int_{-1}^{1} y^l \int_{-1}^{1} x^m |x-y|^{2\delta-1} dxdy \\ &= 2 \sum_{i=0}^{m} \binom{m}{i} \frac{1}{2\delta+i} \sum_{j=0}^{l+m-i} (-1)^j \binom{l+m-i}{j} \frac{2^{2\delta+j+i+1}}{2\delta+j+i+1}, \end{aligned} \tag{33}$$

where $Vc = 2\Gamma(1-2\delta)\sin(\pi\delta)$, $\alpha_l = 0$, odd $(l-k)$ and $T_{l,m} = T_{m,l}$ which is independent of the kernel. In consequence $V(\delta)$ has either both l, m even if k is even, or both odd if k is odd. Therefore $m + l$ is always even.

The bandwidth minimizing the asymptotic MISE in (31) and constant C_{opt} is given by

$$h_A = C_{opt} n^{(2\delta-1)/(2k+1-2\delta)}, \tag{34}$$

$$C_{opt} = C_{opt}(\theta) = \left(\frac{(1-2\delta)}{2k\, I^2(K)} \frac{[k!]^2 V(\theta)}{I(g^{(k)}(x, h_{k,s}))} \right)^{1/(2k+1-2\delta)}, \tag{35}$$

with $I(g^{(k)}) > 0$.

The algorithm proposed by Beran et al. (2016) differences the initial input of the SEMIFAR function for guaranteed stationarity by using a default starting value of $m = 1$. This way one can compute an initial bandwidth for estimating m which is asymptotically

consistent and used in $s = 2$. Plugging $\delta = -0.5$ into (34) yields for $k = 2$ an initial bandwidth of $h_0 = n^{-1/3}$ and gives the lowest possible order of optimal bandwidths for $\delta \in (-0.5, 0.5)$. Then we set $p = p_{max}$ to ensure the model order is chosen only after m and h have been estimated from the data once. Using those first estimates a complete nonparametric trend estimate $\hat{g}(x_t)$ and its level of significance is obtained in iteration $s = 3$ and hence the residuals of Y_t can be calculated for some δ. All its parameters in Z_t are estimated directly from the residuals and maximize the corresponding likelihood function. The optimal model order is chosen following its BIC incorporating all other estimates of the current iteration. The SEMIFAR(p,d,q) fit requires a sufficiently large sample with $\epsilon \sim$ i.i.d $(0, \sigma_\epsilon)$ and the algorithm used in `R` is structured as follows:

Step 1: Search for a bandwidth h_s for the estimation of m^0 and set $s = 1$

- Step 1a: Set $m = 1$ and calculate the FARIMA(p,d,0) process $Y_t = (1 - B)^m Z_t$ described in (8). Estimate g from Y_t in line with (14) using the initial bandwidth $h_{s-1} = n^{-1/3}$ and calculate the residuals.
- Step 1b: Set $p = p_{max}$ assuming the residuals process to follow a FARIMA(p,d,0) model. Set the next bandwidth $h_s = (h_{s-1})^\alpha$ with $\alpha = \hat{\alpha}_{opt} = (5 - 2\hat{\delta})/(7 - 2\hat{\delta})$ and improve h_{s-1} by

$$h_s = C_{opt} \cdot n^{(2\delta-1)/(2k+1-2\delta)}, \tag{36}$$

  with $I(K)$ and estimator $\hat{I}(g^{(k)}(\hat{x}, h_{2,s})$ given in (27) and (26) respectively using the bandwidth $h_{k,s}$. The estimate $\hat{V}$ gives the constant in the asymptotic variance.

Step 2: Set $s = 2$ and estimate m^0

- Step 2a: Carry out Steps 1a and 1b with the selected h_s as the updated initial bandwidth for $m = 0$ and $m = 1$ separately.
- Step 2b: Let the BIC select the appropriate m for the estimate $\hat{m}$ of m^0
- Step 2c: Set $m = \hat{m}$

Step 3: Increase s by 1 and repeat step 1a to step 3 for each p in (p, p_{max}) with $m = \hat{m}$ and updated bandwidth $h_s = n^{-5/7}$ until convergence is reached or a given number of iterations has been reached.

Step 4: Select the best autoregressive order p according to the BIC and take the parameter vector estimate θ corresponding to $\hat{p}$ as the final estimate

This faster algorithm features an exponential inflation method (EIM) to reduce the necessary iterations for an appropriate bandwidth choice. Here according to Beran and Feng (2002a) the rate of convergence of $\hat{h}$ has been improved compared to previous algorithms by the use of α_{opt} and is of order $O_p(n^{2(2\delta-1)/(7-2\delta)})$. Since the formulas connected to the trend estimation by local polynomial smoothing have been presented in derivative form including ν, estimation of the derivatives is analogous to the original trend. (14) allows the estimation of trend derivatives $\hat{g}^{(\nu)}(x)$ with $(\nu > 0)$ and is conducted also with a local polynomial fit in a function called `smooth.lpf`.

The fundamental logic behind the iterative plug-in algorithm is a circular relationship between nonparametric kernel smoothing of the trend estimate using parametric input, and parametric estimation of θ for a fractional ARIMA error. Even though the plug-in values may not be optimal, the current trend and associated bandwidth h_s improve every time the previous bandwidth h_{s-1} is updated. To this end, each iteration h_{s-1} is inflated and improved using the current and better fit of C_{opt} in (35) including the nonparametric trend estimate and the long memory parameter within the fractional ARIMA error. Iterating over s minimizes the MISE in (31) such that $\hat{g}^{(\nu)}$ converges to the true $g^{(\nu)}$. Based on the estimated results one can conclude whether the process is moved by a stationary or difference stationary short memory or long memory component, all possibly including a deterministic trend component.

5 Implementation in R

The implementation of the SEMIFAR model in `R` is called via the `semifar.lpf` function. The script covers the kernel estimation of a nonparametric trend allowing for different error structures and maximum likelihood estimation of the parametric part of the model. The `R` function is designed for a fast and uncomplicated application of SEMIFAR and assumes some not necessarily fixed values a priori to be constant. For instance, the level of confidence used for η is by default set to 0.05 and is among other constants not included as input. Nevertheless the script can be made more flexible very easily by using more parameters in the function call. Besides the series there are currently seven variables to be set before the model is fitted. The minimal autoregressive order p, the maximal order p, the range of the MSE of the trend estimate, the degree of the polynomial approximating

the nonparametric trend and its derivatives, the used kernel, the inflation factor inflating the previously estimated bandwidth and whether boundary protection is included or not.

Fitting a SEMIFAR process with the IPI algorithm required the purchase of the TIBCO Spotfire software `SPLUS` and the additional `FinMetrics` package based on the `S` language. To this date, the `semifar.lpf` function has successfully been implemented into the language R. Major differences in computation are found in

1. the derivation of the ML parameter estimates for a fractionally-differenced ARIMA model, especially for the fractional filter δ
2. how the residuals of fractionally filtered data for an ARIMA model are obtained
3. the calculation of the spectral density and its update
4. a newly introduced inflation factor for bandwidth convergence

Concerning the first point, the difficulty in calculating δ comes with introducing antipersistent trend behavior allowing $\delta < 0$ and extending the interval of possible values for the `drange` to $(-0.5, 0.5)$. The main part of producing δ is given in (2) and implements the gamma function and the backshift operator in an infinte binomial sum over k subtracted from one such that $1 - \delta B - \frac{1}{2}\delta(1-\delta)B^2 - \frac{1}{6}\delta(1-\delta)(2-\delta)^2B^3 - \ldots$ which gives $1 - \sum_{k=1}^{\infty} c_k(\delta)B^k$. For negative δ the signs reverse and the expression is < 0. However, the values in the interval `drange` in `fracdiff{fracdiff}` (over which the likelihood function is maximized as a function of δ) cannot exceed 0.5 since the binomial approaches 0 for $\delta = |\{-0.5, 0.5\}|$ and the series would be non stationary. Still, having δ set to $(-0.5, 0.5)$ is crucial for the identification of the trend mechanism of the SEMIFAR process, as the spectral density $f_X(\lambda) \sim \{0, c, \infty\}$ for δ in $\{(-0.5, 0), 0, (0, 0.5)\}$ in order to obtain the autocovariance behavior at the 0th frequency $f_X(0) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma_X(k) = \{0, c, \infty\}$ expressing the memory structure of the process. Therefore a decision criteria for setting `drange` to $(0, 0.5)$ or to $(-0.5, 0)$ is introduced. If either of the specified intervals for `drange` does not allow the computation of a correlation matrix of the parameter estimates, the alternative `drange` is chosen. In this way, the underlying temporal dependence structure of the data has to correlate positively with the fractional filter. Otherwise there must not be a correlation matrix.

```
1 > m1=fracdiff(data,drange = c(0,0.5))
2 > m2=fracdiff(data,drange = c(-0.5,0))
```

```
> if(isTRUE(all.equal(m1$correlation.dpq,NULL))){result<-m2}
> if(isTRUE(all.equal(m2$correlation.dpq,NULL))){result<-m1}
```

By implementing two different persistence situations, `fracdiff` objects and a decision criterion, the fractional parameter is successfully estimated and equivalent to `arima.fracdiff` output in `S`.

The second change accounts for a different filtering technique used to obtain the residuals and variance of the fractionally filtered ARIMA process. The residuals account for the difference between an estimated and observed value. While in `S` a dynamic linear Kalman filter is employed to calculate the residuals, a much simpler autoregressive moving average filter is used in `R`. The Kalman filter predicts the future state of the series based on the current state vector times the change in time and gives the expected subsequent state. The previous predicted state is corrected using a current sensory measurement as innovation of the state estimate. However, the innovation is weighted relative to the prediction each time step. Its weight decreases if the current measurement matches the last predicted state and increases otherwise. Hence, the ratio of uncertainty in measurement and prediction set by the estimated ARMA parameters, filters for short memory and produces the residuals. An approach more straightforward is the successive application of standard recursive and convolution filtering methods for four different model order- and parameter estimates as filter input for the new `arma.filt` function. In this way all possible ARMA(p,q) contributions to short memory are addressed. Convolution filtering uses the product of different functions as weighted average and thus creates an individually weighted mean for each observation (therefore moving average) in line with

$$I(x,y) = \sum_{s=-i}^{i} \sum_{t=-j}^{j} w(s,t)(x-s, y-t). \tag{37}$$

Hence there are at least $max\{p,q\}$ `NA`s for each filter value. The recursive filtering method describes the re-use of its output as input which is also done within the Kalman filter. The filtered values will exceed their input values for positive w_j and will be lower for

$w_j < 0$. The recursive filter method is given by

$$y_t = \sum_{j=0}^{J} y_{t-j} w_j, \tag{38}$$

where J is the set of filter criteria. Combining both filter mechanics ensures the consistent elimination of the short memory ARMA component moving the input series. Due to the recursive nature of the AR filter, the MA filter obviously has to be applied before the recursive filter. It is important to use the (fractionally) pre-filtered data since otherwise the filter will rule out movement which does not reflect short memory exclusively.

The third difference comes with the calculation of the spectral density update. `S` and `R` produce identical output, apart from the density spectrum which is no longer estimated in decibels such that the transformation `f=10**f$spec/10` is redundant and only `f=f$spec` remains. Moreover the frequency of the input `xfreq` can be fixed to one since we are considering an univariate time series.

The fourth change also concerns the update cycle by allowing different factors α used for inflating the previous bandwidth estimate in (36). The inflation method significantly changes the rate of convergence (Beran and Feng, 2002a). The optimal rate of convergence h_M expresses a trade-off between a bias and variance ratio producing an overall fit of the trend which is sufficiently stable while being sufficiently smooth. The `semifar.lpf` function offers the choice for three different inflation factors called `var`, `naive` and `opt`. If α is chosen as $\alpha_{opt} = (5 - 2\delta)/(7 - 2\delta)$ and δ is for simplicity assumed to be zero, than we have respectively

$$\begin{aligned} \hat{h} &= h_M \left\{1 + O(n^{2(2\delta-1)/(7-2\delta)}) + O_p(n^{2(2\delta-1)/(7-2\delta)}) + O_p(n^{(-1/2)})\right\} \\ &= h_M \left\{1 + O(n^{-2/7}) + O_p(n^{-2/7}) + O_p(n^{-1/2})\right\} \end{aligned} \tag{39}$$

where the first term gives the order of the bias under this particular inflation factor and the second term corresponds to the order of the variance. The third term is asymptotically negligible and for all choices of α the same. The overall rate of convergence under this inflation factor is $O(n^{-2/5})$. The lower bound of the variance is given if $O_p(n^{-1/2})$ since δ is in $(-0.5, 0.5)$ and the rearranged exponent contains $\delta - 1/2$ which is zero if the maximal

value of δ is inserted. If the model includes persistent behavior and δ is non zero the rate of convergence is slower for $\delta > 0$ and respectively for $\delta < 0$ the rate of convergence is faster. Depending on α, either the variance term or the bias is of a smaller order and changes smoothing or stabilizes the trend function more. In general the best overall rate of convergence is achieved by applying an inflation factor α that minimizes the bias and variance of the SEMIFAR model.

6 Applications

Following the spatial definition of a SEMIFAR process in (16), the trend $\hat{g}(x_{t_0}, \tau)$ can be obtained in line with (23). For that purpose the dataset is investigated time-successively ideally for a longer period of time (e.g. 8 years). For computational reasons only each kth value is considered and chosen following the k indexing method, where for each day $m = 511$[1] elements with distance k to the previous element are drawn from the data.

Using the return series as input we can specify a surface plot which visualizes the movement of the returns over the course of the trading day and shows its path across the years along the second horizontal axis. The approach that is necessary to generate figure 1 comes with restructuring the chronological data and its time order. All consecutive 511 trading minutes of the trading day are captured on the x axis, the trading day of 250 days in a year on the y axis and the corresponding return value is given on the z axis. Once the trend in both directions can be calculated we can also receive the trend adjusted residuals or returns. The trend surface for the BMW returns is also given in figure 1. Subtracting the trend produces the desired stationary short (long) memory lattice process without a deterministic or stochastic trend given in 2.

6.1 Over given trading days

Trend investigation requires a time series with sufficiently many observations in order to apply a long memory approach. Here BMW returns between the period of January 2007 and July 2014 are represented. The time-in-the-day dimension is used to illustrate the

[1]The number 511 gives all active trading minutes on a normal 8.5 hours trading day (09:00 - 17:30) of the Deutsche Brse.

trend behavior against the typical pattern of higher trading activity associated to market opening and closing times. Te returns show a peak in then end of 2008 and in 2011 across all time-in-the-day observations. We investigate the trend mechanisms in each direction separately. Three of the almost 2000 trading days of BMW returns are selected and shown in figure 3. Neglecting the absolute values of the returns the density of the trading activity is visibly at lowest levels during 12 am and 2 pm. Higher trading activity expressed as volatility clustering of the return values can be observed towards market opening and closing times. The chosen kth value at midday is apparently more distant to its previous value since fewer trading activities lead to fewer ticks between 12 and 2 pm and therefore reduce the density of the volatility. This phenomenon is referred to as the saddle pattern.

6.2 Over given trading times

The dataset is additionally investigated regarding the trading times over a day such that the first minute of the first trading day is followed by the first minute of all considered trading days, thereafter follows the second minute of the first, second and third day and so on. The restructuring of the data allows trend investigation while simultaneously accounting for the daily saddle pattern. Two of 510 trading minutes were selected and show the BMW returns of 8 years of trading in figure 5.

Notice the two peaks in returns in 2008 during the financial crises and in the year of sales record 2011. Both irregularities have also been shown in figure 1 and indicate a period of high return volatility. The spatial SEMIFAR allows the separation of a trend and stationary short- and long- memory component. By subtracting the trend from the returns

7 Application to high frequency financial data

8 Conclusion/Final remarks

Reviewing the conducted work we conclude, that

References

Angulo, JM, MD Ruiz-Medina, and VV Anh (2000). "Estimation and filtering of fractional generalised random fields". In: *Journal of the Australian Mathematical Society* 69.3, pp. 336–361.

Anh, VV, JM Angulo, and MD Ruiz-Medina (1999). "Possible long-range dependence in fractional random fields". In: *Journal of Statistical Planning and Inference* 80.1-2, pp. 95–110.

Beran, Jan (1995). "Maximum likelihood estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 659–672.

Beran, Jan, RJ Bhansali, and D Ocker (1998). "On unified model selection for stationary and nonstationary short-and long-memory autoregressive processes". In: *Biometrika* 85.4, pp. 921–934.

Beran, Jan and Yuanhua Feng (2002a). "Iterative plug-in algorithms for SEMIFAR modelsdefinition, 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 modelsa 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, Sucharita Ghosh, and Dieter Schell (2009). "On least squares estimation for long-memory lattice processes". In: *Journal of Multivariate Analysis* 100.10, pp. 2178–2194.

Beran, Jan et al. (2016). *Long-Memory Processes.* Springer.

Chen, Chun-Hung, Wei-Choun Yu, and Eric Zivot (2012). "Predicting stock volatility using after-hours information: Evidence from the NASDAQ actively traded stocks". In: *International Journal of Forecasting* 28.2, pp. 366–383.

Cox, DR (1984). "A review". In: *Statistics: An Appraisal, HA David and HT David (Eds.), The Iowa State University Press, Ames, Iowa*, pp. 55–74.

Fan, Jianqing and Irene Gijbels (1995). "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 371–394.

Feng, Y (2004). "Non-and Semiparametric Regression with Fractional Time Series Errors-Theory and Applications to Financial Data". In: *Habilitation Monograph, University of Konstanz*.

Feng, Yuanhua (2003). "Kernel dependent functions in nonparametric regression with fractional time series errors". In:

Gasser, Theo and Hans-Georg Müller (1984). "Estimating regression functions and their derivatives by the kernel method". In: *Scandinavian Journal of Statistics*, pp. 171–185.

Granger, Clive WJ and Roselyne Joyeux (1980). "An introduction to long-memory time series models and fractional differencing". In: *Journal of time series analysis* 1.1, pp. 15–29.

Guo, Hongwen, Chae Young Lim, and Mark M Meerschaert (2009). "Local Whittle estimator for anisotropic random fields". In: *Journal of Multivariate Analysis* 100.5, pp. 993–1028.

Hall, Peter and Jeffrey D Hart (1990). "Nonparametric regression with long-range dependence". In: *Stochastic Processes and Their Applications* 36.2, pp. 339–351.

Hampel, FR (1987). "Data analysis and self-similar processes". In: *Proceedings of the 46th Session of the International Statistical Institute*. International Statistical Institute, pp. 235–254.

Härdle, Wolfgang and Marlene Müller (1997). *Multivariate and semiparametric kernel regression*. Tech. rep. Discussion Papers, Interdisciplinary Research Project 373: Quantification.

Hosking, Jonathan RM (1981). "Fractional differencing". In: *Biometrika* 68.1, pp. 165–176.

Kunsch, Hans Rudolph (1987). "Statistical aspects of self-similar processes". In: *Proceedings of the First World Congress of the Bernoulli Society, 1987*. Vol. 1. VNU Science Press, pp. 67–74.

Lavancier, Frédéric (2007). "Invariance principles for non-isotropic long memory random fields". In: *Statistical inference for stochastic processes* 10.3, pp. 255–282.

Lavancier, Frédéric et al. (2008). "The V/S test of long-range dependence in random fields". In: *Electronic Journal of Statistics* 2, pp. 1373–1390.

Mandelbrot, Benoit B (1983). *The fractal geometry of nature*. Vol. 173. WH freeman New York.

Ruppert, David and Matthew P Wand (1994). "Multivariate locally weighted least squares regression". In: *The annals of statistics*, pp. 1346–1370.

Scott, David W (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Wand, Matt P and M Chris Jones (1994). *Kernel smoothing*. Chapman and Hall/CRC.

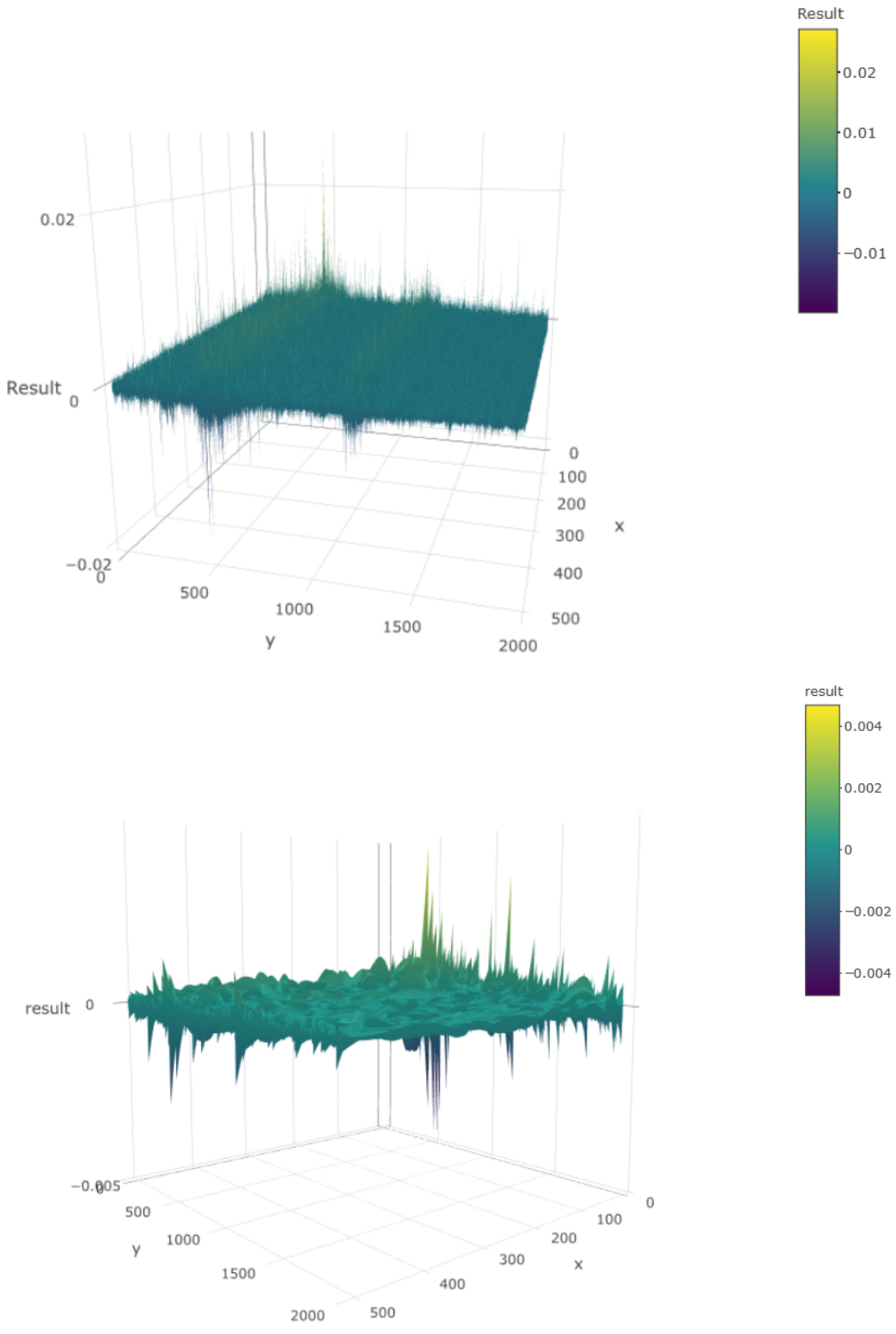


Figure 1: BMW returns and trend 2007 - 2014

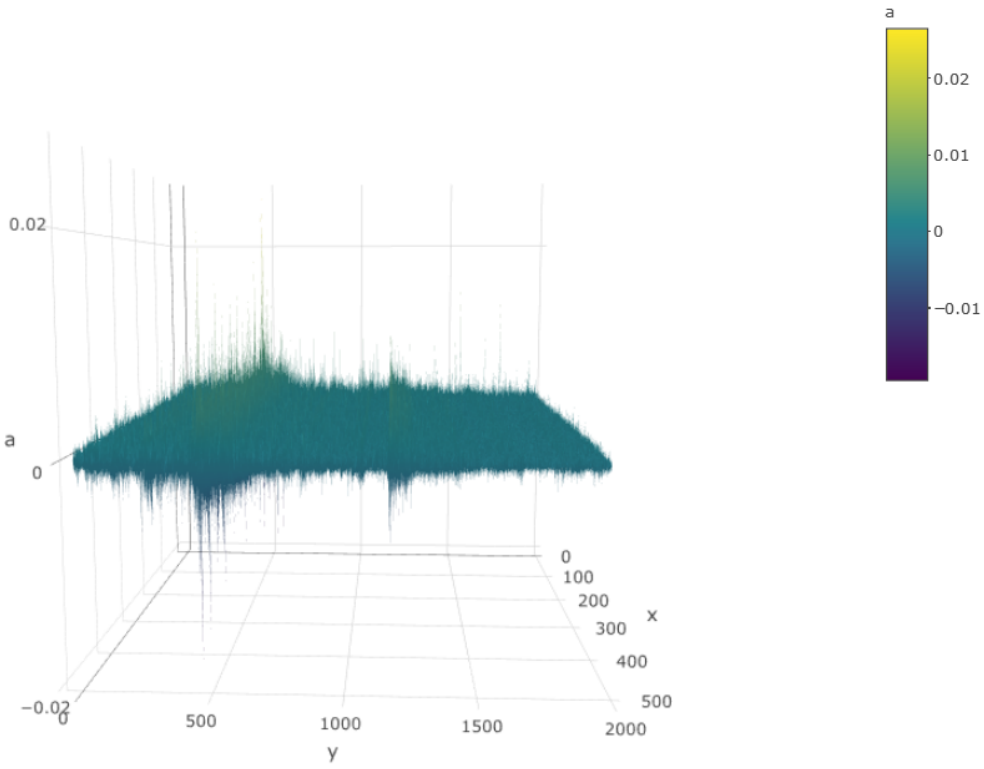


Figure 2: BMW returns trend adjusted

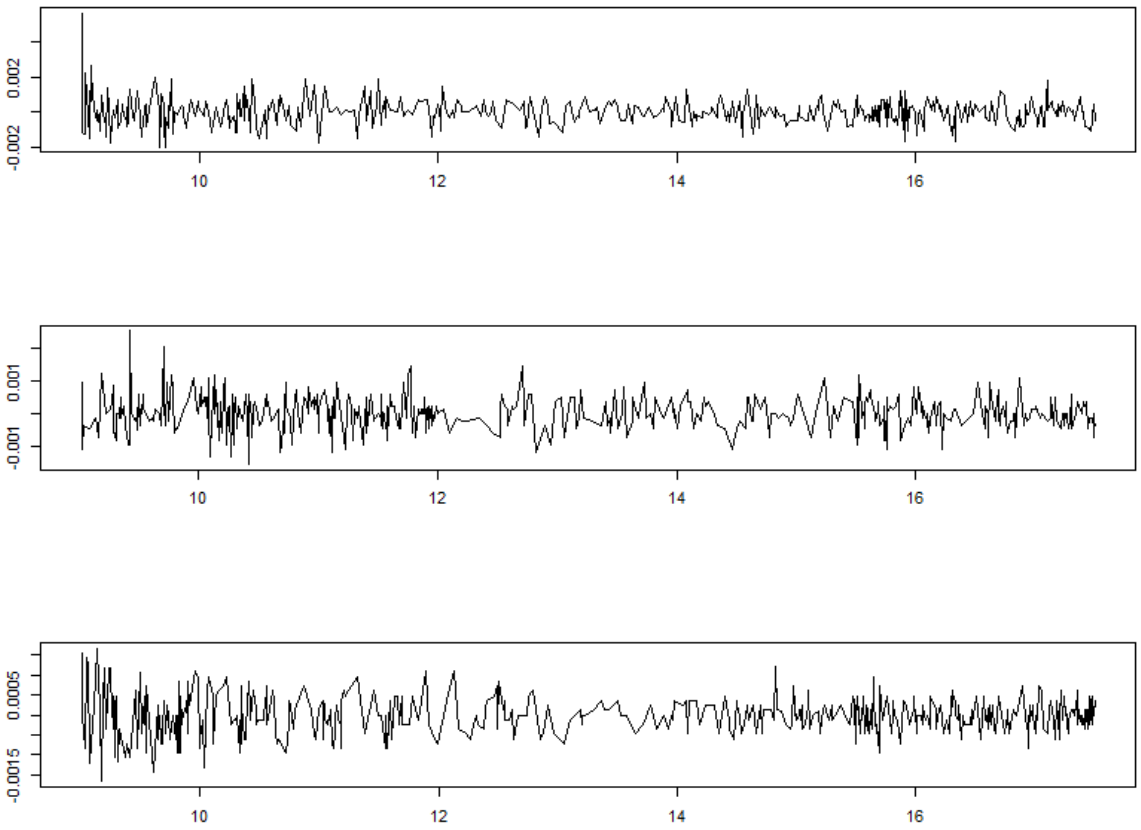
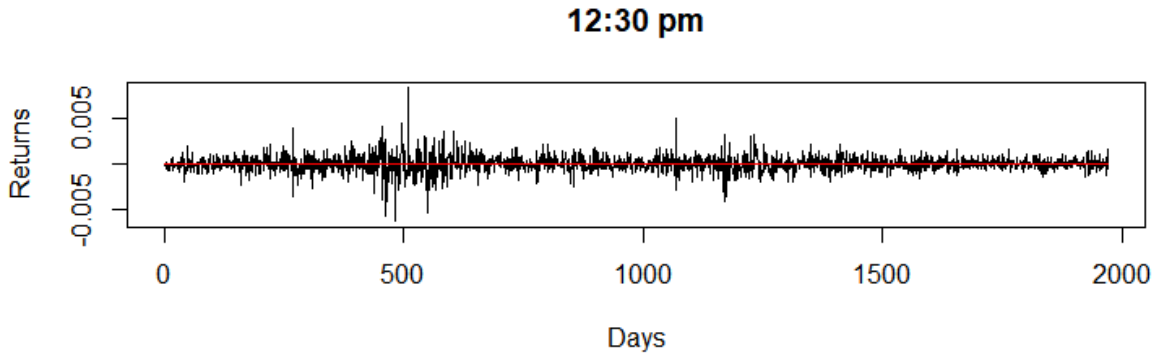


Figure 3: BMW returns trading days 2007 - 2014



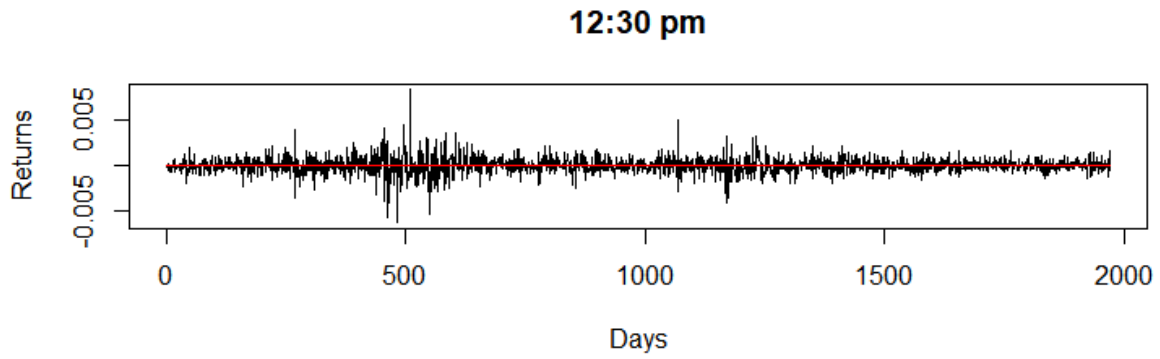


Figure 4: BMW price day time series

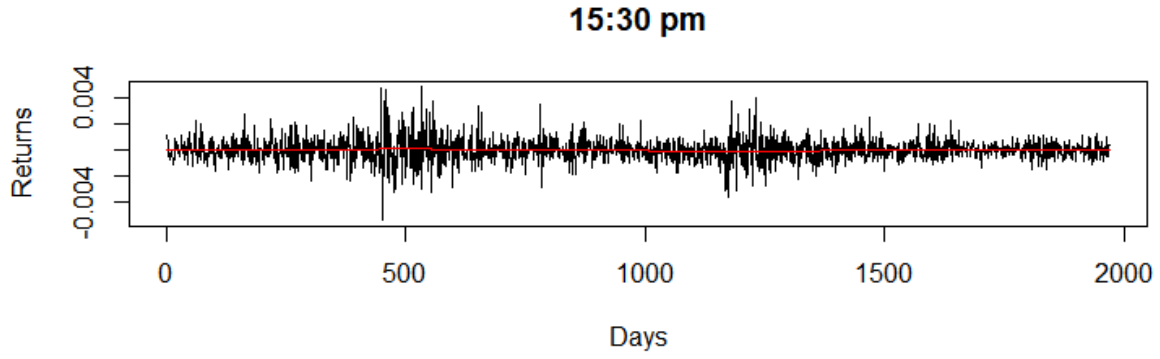


Figure 5: BMW price day time series