Smoothing long memory time series using R

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

December 11, 2020

Abstract

This paper provides first a brief summary of the SEMIFAR (semiparametric fractional autoregressiove) and ESEMIFAR (exponential SEMIFAR) models. Those models are extended slightly to include the moving average part. Under common distribution condition it is shown that the long memory parameter is not affected by the log-transformation. A simple data-driven algorithm is proposed, by which the selected bandwidth and the selected orders of the ARMA model are all consistent. An R package is developed for practical implementation. The application of the proposals are illustrated by different kind of time series.

Keywords: Nonparametric regression with long memory, SEMIFAR, ESEMIFAR, bandwidth selection, model selection, implementation in R,

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 Local polynomial regression with long memory

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 $Y_t = \ln(X_t)$, $g(\tau_t) = \ln[s(\tau_t)]$, $Z_t = \ln(\lambda) + \epsilon_t$ 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

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

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. We have $X_t^* = \exp(Z_t)$. Subsequently, model (2) is an extended version of an ESEMIFAR introduced by Beran, Feng, and Ghosh (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).

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. $\hat{g}^{(\nu)}$ ($\nu \leq l$) 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 $c \in [0, 1)$. Beran and Feng (2002a) and Beran and Feng (2002b) 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 $I[g^{(m)}] = \int_{\Delta}^{1-\Delta} [g^m(\tau)] d\tau$ with Δ being a small positive constant, which controls for the so-called boundary effect. Moreover, $\beta = \int_{-1}^{1} u^m K(u) du$ and for $d > 0$ we have $V(1) = 2c_f \Gamma(1-2d) \sin(\pi d) \int_{-1}^{1} \int_{-1}^{1} K(x) K(y) |x-y|^{2d-1} dx dy$. 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

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

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.

3 Data-driven estimation

3.1 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:

Step 1: In the first iteration set the initial bandwidth h_0 and select p and q denoting the AR- and MA-order, respectively.

Step 1a): Estimate g from Y_t employing h_{j-1} in order to calculate the residuals $\tilde{Z}_t = Y_t - \hat{g}(\tau_t)$.

Step 1b): Obtain $\hat{c}_f$ by fitting a FARIMA (with predefined AR- and MA-order in Step 1) to $\tilde{Z}_t$.

Step 1c): Set $h_j = (h_{j-1})^{\alpha}$, where $\alpha =$ denotes an inflation factor and estimate $I[g^{(m)}]$ with a local polynomial of order $r^* = r + 2$. Now, we obtain h_{j-1} by

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

Step 2: Repeat steps 1a to 1c until convergence or a given number of iterations has been reached and set $h_{opt} = h_j$.

Please note that the results presented in Beran and Ocker (1999), Beran and Ocker (2001), Beran and Feng (2002a), Beran and Feng (2002b), Beran and Feng (2002c) and Beran et al. (2016) remain valid for the IPI for SEMIFARIMA models.

4 Implementation in R

5 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), *gdpUS* (US

GDP), *dax* (German stock index) and *vix* (CBOES volatility index) Those data sets were already used by Feng et al. (forthcoming) and are implemented in the *smoots* package, which was recently published on the *CRAN* network.

5.1 Application of the SEMIFARIMA

5.2 Application of the ESEMIFARIMA

5.3 Application to high-frequency financial data

6 The Semi-FI-Log-GARCH model

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. An extension of this model is the Log-GARCH introduced by Pantula (1986), Geweke (1986) and Milhøj (1987). The majority of the GARCH class models are defined under the assumptions that the considered return series is stationary. and incorporates a short-memory dependence structure. However, it was found that in practice return series exhibit a slowly changing scale and long term dynamics in the conditional volatility.

7 The Semi-FI-Log-ACD model* (nachfragen)

8 Concluding remarks

References

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.

Beran, Jan and Yuanhua Feng (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.

– (2001). "Volatility of stock-market indexes—an analysis based on SEMIFAR models". In: *Journal of Business & Economic Statistics* 19.1, pp. 103–116.

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.

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.

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 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.

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.

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

Appendix