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 The SEMIFARIMA model

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 $E(\epsilon_t) = 0$, $\mathrm{var}(\epsilon_t) = \sigma_\epsilon^2$, $x_t = t/n$ 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 $\gamma(k) = \mathrm{cov}(Y_t, Y_{t+k})$, 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 IPI (iterative plug-in) algorithm already introduced in Beran (1995) by replacing an estimate of the constant mean with a kernel estimate of g defined by

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

where $h > 0$ denotes the bandwidth, $x \in [0,1]$ and $K(\cdot)$ is a symmetric second order kernel with compact support (see e.g. Gasser and Müller 1979). Moreover, explicit expressions for the bias, variance, MISE and the optimal bandwidth which minimises the asymptotic MISE are stated in Theorem 1 in Beran and Feng (2002c). A comprehensive application of the SEMIFAR to finanical time series data was carried out by Beran and Ocker (2001). The authors found strong evidence of long memory in power transformed absolute return series in form of a stochastic- or deterministic trend and in some cases with both forms. Subsequently, these results indicate that conventional parametric short- and long memory models may not be suitable for modelling volatility of stock market indices. In Beran and Feng (2002a) two new IPI-algorithms are proposed which run much faster as they do not rely on a full search of the long memory parameter. Furthermore, an EIM- (exponential inflation method) bandwidth selector is defined. Beran and Ocker (2001) and Beran and Feng (2002c) already suggested to use an EIM as it requires less iterations than the conventional multiplicative inflation method (MIM) used by Gasser et al. (1991), Herrmann et al. (1992) and Ray and Tsay (1997). Different choices for the inflation factor and asymptotic properties of the estimated bandwidths are derived by Beran and Feng (2002a). To control for the poor estimation quality of the kernel estimator at the boundaries, the authors introduced a small positive constant $\Sigma > 0$ such that as $n \to \infty$, $h \to 0$, $nh \to \infty$,

$$\text{MISE} = E\Big\{ \int_{\Sigma}^{1-\Sigma} [\hat{g}(x) - g(x)]^2 dx \Big\}. \tag{7}$$

In a following paper Beran and Feng (2002b) replaced the kernel estimator (6) with a local local approximation of $g(x)$ given by the p order polynomial

$$g(x_t) \approx g(x) + g^{(1)}(x)(x_t - x) + \ldots + g^{(p)}(x)\frac{(x_t - x)^p}{p!} + R_p, \tag{8}$$

where R_p is a remainder term. 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

$$Q(x) = \sum_{t=1}^{n} \left\{ Y_t - \sum_{j=0}^{p} \beta_j (x_t - x)^j \right\}^2 K\left(\frac{x_t - x}{h}\right) \Rightarrow min, \tag{9}$$

where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ and h and K are defined as before.

3 The E-SEMIFARIMA model

A theorem about the memory property of the ESEMIFAR (direct after that model)

Theorem 1. *Assume that X_t follows an ESEMIFAR model defined above with $d > 0$ and a linear FARIMA error process ξ_t in the log-data, then $Z_t = X_t / v(\tau_t)$ is a stationary long memory process with the same memory parameter d.*

This result shows that the level of long memory of the stationary part of the process under consideration is not affected by the log- or exponential transformations. In the case when ξ_t is normal, those results are e.g. obtained in Dittmann and Granger (2002) and Beran et al. (2015). Theorem 2 extends those results to a common innovation distribution in the log-data. The proof of Theorem 2 is given in the appendix, where we will show that the process Z_t is a special case of the general framework defined in Surgailis and Viano (2002). Hence, their results apply to Z_t. Note however that the above result does not hold for $d = 0$. Detailed discussion on corresponding properties in case with $d = 0$ is beyond the aim of the current paper.

4 Data-driven estimation

Theorem 2. *Under assumptions A1 - A5, the selected AR order $\hat{p}_1$ and MA order $\hat{q}_1$ at the end of the first out-side iteration are both consistent.*

The proof of Theorem 1 is given in the appendix. Note in particular that to achieve consistent model selection only the first out-side iteration is required. Hence, if p_0 and q_0 happen to be chosen correctly, the procedure will usually be stopped after the first out-side iteration. Otherwise, it will usually be stopped after the second out-side iteration. If n is large enough, in most of the cases the third or further out-side iterations are not required. This ensures that the running time of the proposed algorithm comparable to that required by the original SEMIFAR algorithm.

5 Implemenation in R

6 Application to different kinds of time series

6.1 Application of the SEMIFARIMA

6.2 Application of the ESEMIFARIMA

6.3 Application to high-frequency financial data

7 Concluding remarks

Appendix